Economic Policy Institute

Report | January 28, 2013

WHAT DO INTERNATIONAL TESTS REALLY SHOW ABOUT U.S. STUDENT PERFORMANCE?

BY MARTIN CARNOY, STANFORD GRADUATE SCHOOL OF EDUCATION AND EPI
AND RICHARD ROTHSTEIN, EPI

Executive summary

Education policymakers and analysts express great concern about the performance of U.S. students on international tests. Education reformers frequently invoke the relatively poor performance of U.S. students to justify school policy changes.

In December 2012, the International Association for the Evaluation of Educational Achievement (IEA) released national average results from the 2011 administration of the Trends in International Mathematics and Science Study (TIMSS). U.S. Secretary of Education Arne Duncan promptly issued a press release calling the results "unacceptable," saying that they "underscore the urgency of accelerating achievement in secondary school and the need to close large and persistent achievement gaps," and calling particular attention to the fact that the 8th-grade scores in mathematics for U.S. students failed to improve since the previous administration of the TIMSS.

Two years earlier, the Organization for Economic Cooperation and Development (OECD) released results from another international test, the 2009 administration of the Program for International Student Assessment (PISA). Secretary Duncan's statement was similar. The results, he said, "show that American students are poorly prepared to compete in today's knowledge economy. … Americans need to wake up to this educational reality—instead of napping at the wheel while emerging competitors prepare their students for economic leadership." In particular, Duncan stressed results for disadvantaged U.S. students: "As disturbing as these national trends are for America, enormous achievement gaps among black and Hispanic students portend even more trouble for the U.S. in the years ahead."

However, conclusions like these, which are often drawn from international test comparisons, are oversimplified, frequently exaggerated, and misleading. They ignore the complexity of test results and may lead policymakers to pursue inappropriate and even harmful reforms.

TABLE OF CONTENTS

www.epi.org

Both TIMSS and PISA eventually released not only the average national scores on their tests but also a rich international database from which analysts can disaggregate test scores by students' social and economic characteristics, their school composition, and other informative criteria. Such analysis can lead to very different and more nuanced conclusions than those suggested from average

national scores alone. For some reason, however, although TIMSS released its average national results in December, it scheduled release of the international database for five weeks later. This puzzling strategy ensured that policymakers and commentators would draw quick and perhaps misleading interpretations from the results. This is especially the case because analysis of the international database takes time, and headlines from the initial release are likely to be sealed in conventional wisdom by the time scholars have had the opportunity to complete a careful study.

While we await the release of the TIMSS international database, this report describes a detailed analysis we have conducted of the 2009 PISA database. It offers a different picture of the 2009 PISA results than the one suggested by Secretary Duncan's reaction to the average national scores of the United States and other nations.

Because of the complexity and size of the PISA international database, this report's analysis is restricted to the comparative test performance of adolescents in the United States, in three top-scoring countries, and in three other post-industrial countries similar to the United States. These countries are illustrative of those with which the United States is usually compared. We compare the performance of adolescents in these seven countries who have similar social class characteristics. We compare performance in the most recent test for which data are available, as well as trends in performance over the last nearly two decades.

In general, we find that test data are too complex and oversimplified to permit meaningful policy conclusions regarding U.S. educational performance without deeper study of test results and methodology. However, a clear set of findings stands out and is supported by all data we have available:

Because social class inequality is greater in the United States than in any of the countries with which we can reasonably be compared, the relative performance of U.S. adolescents is better than it appears when countries' national average performance is conventionally compared.

- Because in every country, students at the bottom of the social class distribution perform worse than students higher in that distribution, U.S. average performance appears to be relatively low partly because we have so many more test takers from the bottom of the social class distribution.
- A sampling error in the U.S. administration of the most recent international (PISA) test resulted in students from the most disadvantaged schools being over-represented in the overall U.S. test-taker sample. This error further depressed the reported average U.S. test score.
- If U.S. adolescents had a social class distribution that was similar to the distribution in countries to which the United States is frequently compared, average reading scores in the United States would be higher than average reading scores in the similar post-industrial countries we examined (France, Germany, and the United Kingdom), and average math scores in the United States would be about the same as average math scores in similar post-industrial countries.
- A re-estimated U.S. average PISA score that adjusted for a student population in the United States that is more disadvantaged than populations in otherwise similar post-industrial countries, and for the oversampling of students from the most-disadvantaged schools in a recent U.S. international assessment sample, finds that the U.S. average score in both reading and mathematics would be higher than official reports indicate (in the case of mathematics, substantially higher).
- This re-estimate would also improve the U.S. place in the international ranking of all OECD countries, bringing the U.S. average score to sixth in reading and 13th in math. Conventional ranking reports based

on PISA, which make no adjustments for social class composition or for sampling errors, and which rank countries irrespective of whether score differences are large enough to be meaningful, report that the U.S. average score is 14th in reading and 25th in math.

- Disadvantaged and lower-middle-class U.S. students perform better (and in most cases, substantially better) than comparable students in similar post-industrial countries in reading. In math, disadvantaged and lower-middle-class U.S. students perform about the same as comparable students in similar post-industrial countries.
- At all points in the social class distribution, U.S. students perform worse, and in many cases substantially worse, than students in a group of top-scoring countries (Canada, Finland, and Korea). Although controlling for social class distribution would narrow the difference in average scores between these countries and the United States, it would not eliminate it.
- U.S. students from disadvantaged social class backgrounds perform better relative to their social class peers in the three similar post-industrial countries than advantaged U.S. students perform relative to their social class peers. But U.S. students from advantaged social class backgrounds perform better relative to their social class peers in the top-scoring countries of Finland and Canada than disadvantaged U.S. students perform relative to their social class peers.
- On average, and for almost every social class group, U.S. students do relatively better in reading than in math, compared to students in both the top-scoring and the similar post-industrial countries.

Because not only educational effectiveness but also countries' social class composition changes over time, comparisons of test score trends over time by social class group provide more useful information to policymakers than comparisons of total average test scores at one point in time or even of changes in total average test scores over time.

- The performance of the lowest social class U.S. students has been improving over time, while the performance of such students in both top-scoring and similar post-industrial countries has been falling.
- Over time, in some middle and advantaged social class groups where U.S. performance has not improved, comparable social class groups in some top-scoring and similar post-industrial countries have had declines in performance.

Performance levels and trends in Germany are an exception to the trends just described. Average math scores in Germany would still be higher than average U.S. math scores, even after standardizing for a similar social class distribution. Although the performance of disadvantaged students in the two countries is about the same, lower-middle-class students in Germany perform substantially better than comparable social class U.S. students. Over time, scores of German adolescents from all social class groups have been improving, and at a faster rate than U.S. improvement, even for social class groups and subjects where U.S. performance has also been improving. But the causes of German improvement (concentrated among immigrants and perhaps also attributable to East and West German integration) may be idiosyncratic, and without lessons for other countries or predictive of the future. Whether German rates of improvement can be sustained to the point where that country's scores by social class group uniformly exceed those of the United States remains to be seen. As of 2009, this was not the case.

Great policy attention in recent years has been focused on the high average performance of adolescents in Finland. This attention may be justified, because both math and reading scores in Finland are higher for every social class group than in the United States. However, Finland's scores have been falling for the most disadvantaged students while U.S. scores have been improving for similar

social class students. This should lead to greater caution in applying presumed lessons from Finland. At first glance, it may seem that the decline in scores of disadvantaged students in Finland results in part from a recent influx of lower-class immigrants. However, average scores for *all* social class groups have been falling in Finland, and the gap in scores between Finland and the United States has narrowed in each social class group. Further, during the same period in which scores for the lowest social class group have declined, the share of all Finnish students in this group has also declined, which should have made the national challenge of educating the lowest social class students more manageable, so immigration is unlikely to provide much of the explanation for declining performance.

Although this report's primary focus is on reading and mathematics performance on PISA, it also examines mathematics test score performance in earlier administrations of the TIMSS. Where relevant, we also discuss what can already be learned from the limited information now available from the 2011 TIMSS. To help with the interpretation of these PISA and TIMSS data, we also explore reading and mathematics performance on two forms of the U.S. domestic National Assessment of Educational Progress (NAEP).

Relevant complexities are too often ignored when policymakers draw conclusions from international comparisons. Different international tests yield different rankings among countries and over time. PISA, TIMSS, and NAEP all purport to reflect the achievement of adolescents in mathematics (and PISA and NAEP in reading), yet results on different tests can vary greatly—in the most extreme cases, countries' scores can go up on one test and down on another that purport to assess the same students in the same subject matter—and scholars have not investigated what causes such discrepancies. These differences can be caused by the content of the tests themselves (for example, differences in the specific skills that test makers consider to represent adolescent "mathematics") or by flaws in sampling and test administration. Because these differences are revealed in the most cursory examination of test results, policymakers should exercise greater caution in drawing policy conclusions from international score comparisons.

To arrive at our conclusions, we made a number of explicit and transparent methodological decisions that reflect our best judgment. Three are of importance: our definition of social class groups, our selection of comparison countries, and our determination of when differences in test scores are meaningful.

There is no clear way to divide test takers from different countries into social class groups that reflect comparable social background characteristics relevant to academic performance. For this report, we chose differences in the number of books in adolescents' homes to distinguish them by social class group; we consider that children in different countries have similar social class backgrounds if their homes have similar numbers of books. We think that this indicator of household literacy is plausibly relevant to student academic performance, and it has been used frequently for this purpose by social scientists. We show in a technical appendix that supplementing it with other plausible measures (mother's educational level, and an index of "economic, social, and cultural status" created by PISA's statisticians) does not provide better estimates. Also influencing our decision is that the number of books in the home is a social class measure common to both PISA and TIMSS, so its use permits us to explore longer trend lines and more international comparisons. As noted, however, data on these background characteristics were not released along with the national average scores on the 2011 TIMSS, and so our information on the performance of students from different social class groups on TIMSS must end with the previous, 2007, test administration.

In this report, we focus particularly on comparisons of U.S. performance in math and reading in PISA with performance in three "top-scoring countries" (Canada, Finland, and Korea) whose average scores are generally higher

than U.S. scores, and with performance in three "similar post-industrial countries" (France, Germany, and the United Kingdom) whose scores are generally similar to those of the United States. We employed no sophisticated statistical methodology to identify these six comparison countries. Assembling and disaggregating data for this report was time consuming, and we were not able to consider additional countries. We think our choices include countries to which the United States is commonly compared, and we are reasonably confident that adding other countries would not appreciably change our conclusions. If other scholars wish to develop data for other countries, we would gladly offer them methodological advice.

Technical reports on test scores typically distinguish differences that are "significant" from those that are not. But this distinction is not always useful for policy purposes and is frequently misunderstood by policymakers. To a technical expert, a score difference can be miniscule but still "significant" if it can be reproduced 95 percent of the time when a comparison is repeated. But miniscule score differences should be of little interest to policymakers. In general, social scientists consider an intervention to be worthwhile if it improves a median subject's performance enough to be superior to the performance of about 57 percent or more of all subjects prior to the intervention. Such an intervention should be considered "significant" for policy purposes, but, to avoid confusion, we avoid the term "significant" altogether. Instead, for PISA, we consider countries' (or social class groups') average scores to be "about the same" if they are less than 8 test scale points different (even if this small difference would be repeated in 95 of 100 test administrations), to be "better" or "worse" if they are at least 8 but less than 18 scale points different, and "substantially better" or "substantially worse" if they differ by 18 scale points or more. Eighteen scale points in most cases is approximately equivalent to the difference social scientists generally consider to be the minimum result of a worthwhile intervention (an effect size of about 0.2 standard deviations). The TIMSS scale is slightly different from the PISA scale; for TIMSS, the cut points used in this report are 7 and 17 rather than 8 and 18.

With regard to these and other methodological decisions we have made, scholars and policymakers may choose different approaches. We are only certain of this: To make judgments only on the basis of statistically significant differences in national average scores, on only one test, at only one point in time, without regard to social class context or curricular or population sampling methodologies, is the worst possible choice. But, unfortunately, this is how most policymakers and analysts approach the field.

The most recent test for which an international database is presently available is PISA, administered in 2009. As noted, the database for TIMSS 2011 is scheduled for release later this month (January 2013). In December 2013, PISA will announce results and make data available from its 2012 test administration. Scholars will then be able to dig into TIMSS 2011 and PISA 2012 databases and place the publicly promoted average national results in proper context. The analyses that follow in this report should caution policymakers to await understanding of this context before drawing conclusions about lessons from TIMSS or PISA assessments. We plan to conduct our own analyses of these data when they become available, and publish supplements to this report as soon as it is practical to do so, given the care that should be taken with these complex databases.

Part I. Introduction

A 2009 international test of reading and math showed that American 15-year-olds perform more poorly, on average, than 15-year-olds in many other countries. This finding, from the Program for International Student Assessment (PISA),[1] is consistent with previous PISA results, as well as with results from another international assessment of 8th-graders, the Trends in International Mathematics and Science Survey (TIMSS).[2]

From such tests, many journalists and policymakers have concluded that American student achievement lags woefully behind that in many comparable industrialized nations, that this shortcoming threatens the nation's economic future, and that these test results therefore suggest an urgent need for radical school reform.

Upon release of the 2011 TIMSS results, for example, U.S. Secretary of Education Arne Duncan called them "unacceptable," saying that they "underscore the urgency of accelerating achievement in secondary school and the need to close large and persistent achievement gaps" (Duncan 2012). Two years before, upon release of 2009 PISA scores, Duncan said that "…the 2009 PISA results show that American students are poorly prepared to compete in today's knowledge economy. … Americans need to wake up to this educational reality—instead of napping at the wheel while emerging competitors prepare their students for economic leadership." In particular, Duncan stressed the PISA results for disadvantaged U.S. students: "As disturbing as these national trends are for America, enormous achievement gaps among black and Hispanic students portend even more trouble for the U.S. in the years ahead. Last year, McKinsey & Company released an analysis which concluded that America's failure to close achievement gaps had imposed—and here I quote—'the economic equivalent of a permanent national recession.'" The PISA results, Duncan concluded, justify the reform policies he has been pursuing: "I was struck by the convergence between the practices of high-performing countries and many of the reforms that state and local leaders have pursued in the last two years" (Duncan 2010).

This conclusion, however, is oversimplified, exaggerated, and misleading. It ignores the complexity of the content of test results and may well be leading policymakers to pursue inappropriate and even harmful reforms that change aspects of the U.S. education system that may be working well and neglect aspects that may be working poorly.

For example, as Secretary Duncan said, U.S. educational reform policy is motivated by a belief that the U.S. educational system is particularly failing disadvantaged children. Yet an analysis of international test score levels and trends shows that in important ways disadvantaged U.S. children perform better, relative to children in comparable nations, than do middle-class and advantaged children. More careful analysis of these levels and trends may lead policymakers to reconsider their assumption that almost all improvement efforts should be directed to the education of disadvantaged children and few such efforts to the education of middle-class and advantaged children.

Education analysts in the United States pay close attention to the level and trends of test scores disaggregated by socioeconomic groupings. Indeed, a central element of U.S. domestic education policy is the requirement that average scores be reported separately for racial and ethnic groups and for children who are from families whose incomes are low enough to qualify for the subsidized lunch program. We understand that a school with high proportions of disadvantaged children may be able to produce great "value-added" for its pupils, although its average test score levels may be low. It would be foolish to fail to apply this same understanding to comparisons of international test scores.

Extensive educational research in the United States has demonstrated that students' family and community characteristics powerfully influence their school performance. Children whose parents read to them at home, whose health is good and can attend school regularly, who do not live in fear of crime and violence, who enjoy stable housing and continuous school attendance, whose parents' regular employment creates security, who are exposed to museums, libraries, music and art lessons, who travel outside their immediate neighborhoods, and who are surrounded by adults who model high educational achievement and attainment will, on average, achieve at higher levels than children without these educationally relevant advantages. We know much less about the extent to which

similar factors affect achievement in other countries, but we should assume, in the absence of evidence to the contrary, that they do.

It is also the case that countries' educational effectiveness and their social class composition change over time. Consequently, comparisons of test score trends over time by social class group provide more useful information to policymakers than comparisons of total average test scores at one point in time or even of changes in total average test scores over time.

Unfortunately, our conversation about international test score comparisons has ignored such questions. It would be foolish, for example, to let international comparisons motivate radical changes in educational policies in a country whose social class subgroup average scores were below those of other nations, if that country's subgroups had been improving their performance at a more rapid rate than similar subgroups in other nations, even if the country's overall average still had not caught up. Just as a domestic U.S. school's average performance is influenced by its social class composition, so too might a country's average performance be influenced by its social class composition.

The policy responses of educational reformers should be sufficiently nuanced to respond to such considerations, because policy initiatives might improve in response to more sophisticated inquiries.

For example, consider Country C. Its affluent students achieve better than affluent students in comparable countries, but not as much better as in the past; the performance of affluent students in Country C, while still relatively high, has been declining relative to the performance of affluent students in comparable countries. Country C's socioeconomically disadvantaged students achieve less than disadvantaged children in comparable countries, but not as much less as in the past. The performance of disadvantaged students in Country C, while still relatively low, has been improving relative to the performance of disadvantaged students in comparable nations. In such circumstances, unsophisticated reformers in Country C might well decide to revamp how disadvantaged students are being taught, even though teaching methods have been successfully raising such students' achievement relative to the achievement of similarly disadvantaged students in other countries and relative to the achievement of wealthier students in Country C itself. Such unsophisticated reformers might also ignore the condition of education of affluent students, believing that their relatively high performance suggests that no reform is needed, while overlooking the decline of such performance over time. Sophisticated education policymakers, in contrast, who have studied the data trends, might direct their reform efforts to the high-scoring rather than the low-scoring students.

Thus, in evaluating a country's educational performance, we should want to know how children from different social class groups perform, in comparison to other social class groups within their own country and in comparison to children from similar social class groups in other countries. Describing only an "average" national score obscures what is likely to be more useful information. Yet it is only in terms of national averages that policy discussion of international test scores typically proceeds. U.S. policymakers would learn more if they also studied the performances of demographic (socioeconomic) subgroups and compared these to the performances of similar subgroups in other nations. To the extent international comparisons are important, it is critical to know whether each subgroup in the United States performs above or below the level of socioeconomically similar subgroups in comparable industrialized nations.

If we identify subgroups that perform relatively well or relatively poorly in one country or another, we should also ask how the performances of these subgroups, compared to the performances of similar subgroups in other nations, are changing over time. Are some subgroups improving their performance unusually rapidly, in comparison to socioeconomically similar subgroups in other nations,

while other subgroups are exhibiting unusual deterioration in performance? Are various subgroups improving or declining in performance at different rates, and are these differences masked when we look only at national averages?

In this report, we also identify inconsistencies between various international tests that may well be related to inaccurate population sampling that has caused some tests to oversample some social class groups and undersample others. Such sampling errors inevitably lead to inaccuracies in reports of how students in a particular country perform, relative to those in other countries where the sampling may have been more accurate.

Other considerations, rarely considered in public debate, also influence the care we should take in the interpretation of international comparisons. One is how the curriculum is sampled in the framework for any particular test. Because the full range of knowledge and skills that we describe as "mathematics" cannot possibly be covered in a single brief test, policymakers should also carefully examine whether an assessment called a "mathematics" test necessarily covers knowledge and skills similar to those covered by other assessments also called "mathematics" tests, and whether performance on these different assessments can reasonably be compared. For example, American adolescents perform relatively well on algebra questions, and relatively poorly on geometry questions, compared to adolescents in other countries. Reports on how the United States compares to other countries show the United States in a more favorable light to the extent a test has more algebra items and fewer geometry items. Whether there is an appropriate balance between these topics on any particular international assessment is rarely considered by policymakers who draw conclusions about the relative performance of U.S. students from that assessment. Similar questions arise with regard to a "reading" test.

Whether U.S. policymakers want to reorient the curriculum to place more emphasis on geometry is a decision they should make without regard to whether such reorientation might influence comparative scores on an international test. It certainly might not be good public policy to reduce curricular emphasis on statistics and probability, skills essential to an educated citizenry in a democracy, in order to make more time available for geometry. There are undoubtedly other sub-skills covered by international reading and math tests on which some countries are relatively stronger and others are relatively weaker. Investigation of these differences should be undertaken before drawing policy conclusions from international test scores.

To stimulate an examination and discussion of these and several other complexities, we analyze data on the performance of adolescents from PISA and TIMSS, as well as from two forms of the National Assessment of Educational Progress (NAEP), a test given exclusively to a sample of U.S. students. The first form, Main NAEP, is modified in small ways over time, so that its coverage tracks modifications in the math curriculum. The second form, Long-Term Trend NAEP (LTT), which changes much less over time, assesses how students' competence changes over time on a more nearly identical set of skills. The Main NAEP has been administered since 1990, and the LTT since the early 1970s.[3]

Part II. PISA 2009—the comparative performance of U.S. students by social class group

Disaggregation of PISA test scores by social class group reveals some patterns that many education policymakers will find surprising. Average U.S. test scores are lower than average scores in countries to which the United States is frequently compared, in part because the share of disadvantaged students in the overall national population is greater in the United States than in comparison countries. If the social class distribution of the United States were similar to that of top-scoring countries, the average test score gap between the United States and these top-scoring countries would be cut in half in reading and by one-third in mathematics. Dis-

advantaged U.S. students perform comparatively better than do disadvantaged students in important comparison countries. The test score gap between advantaged and disadvantaged students in the United States is smaller than the gap in similar post-industrial countries; it is generally, although not always, greater than the gap in top-scoring countries. This section explores these findings in greater detail.

To simplify our comparisons of national average PISA scores and of these scores disaggregated by social class, we focus on the United States and six other countries—Canada, Finland, South Korea (hereinafter simply Korea), France, Germany, and the United Kingdom.

We refer to three of these countries (Canada, Finland, and Korea) as "top-scoring countries" because they score much better overall than the United States in reading and math—about a third of a standard deviation better.[4] Canada, Finland, and Korea are also the three "consistent high-performers" that U.S. Secretary of Education Arne Duncan highlighted when he released the U.S. PISA results (Duncan 2010).

We call the other three (France, Germany, and the United Kingdom) "similar post-industrial countries" because they score similarly overall to the United States. They also are countries whose firms are major competitors of U.S. firms in the production of higher-end manufactured goods and services for world markets. Their firms are not the only competitors of U.S. firms, but if the educational preparation of young workers is a factor in national firms' competitiveness, it is worth comparing student performance in these countries with student performance in the United States to see if these countries' educational systems, so different from that in the United States, play a role in their firms' success.

PISA is scored on a scale that covers very wide ranges of ability in math and reading. When scales were created for reading in 2000 and for math in 2003, the mean for all test takers from countries in the Organization for Economic Cooperation and Development (OECD), the sponsor of PISA, was set at 500 with a standard deviation of 100. When statisticians describe score comparisons, they generally talk about differences that are "significant." Yet while "significance" is a useful term for technical discussion, it can be misleading for policy purposes, because a difference can be statistically significant but too small to influence policy. Therefore, in this report, we avoid describing differences in terms of statistical significance. Instead, we use terms like "better (or worse)" and "substantially better (or worse)" (both of which are significantly better for statistical purposes), and "about the same."[5]

In general, in this report, we use the term "about the same" to describe average score differences in PISA that are less than 8 scale points; we use the term "better (or worse)" to describe differences that are at least 8 points but less than 18 scale points, and we use the term "substantially (or much) better (or worse)" to describe differences that are 18 scale points or more.[6] Of course, any fixed cut point is arbitrary, and readers may find it strange when we say, for example, that when two countries have an average difference of 7 scale points they perform about the same, whereas when their average difference is 8 scale points one performs better than the other. This is a necessary consequence of any descriptive system using cut points. However, this caution is in order: Readers without statistical sophistication will be tempted to think that a difference of 7 scale points is almost "better." This is true. But a difference of 8 scale points is also almost "about the same." Many readers, accustomed to finding differences where there are none, will be more reluctant to consider the latter than the former, but both are equally true.

Table 1 displays overall average scores in reading and math reported by PISA for 2009. These are the basis (without any socioeconomic disaggregation) of most commonplace comparisons.

The table shows that, on average, U.S. performance was substantially worse than performance in the top-scoring

TABLE 1

Overall average national scale scores, reading and math, for U.S. and six comparison countries, PISA 2009

	TOP SCORING				SIMILAR POST-INDUSTRIAL				U.S.	U.S. VERSUS:	
	Canada	Finland	Korea	Average*	France	Germany	U.K.	Average*		Top-scoring average	Similar post-industrial average
Reading	524	536	539	533	496	497	494	496	500	-33	+4
Math	527	541	546	538	497	513	492	501	487	-50	-13

* Simple (unweighted) average of three countries

Source: Authors' analysis of OECD Program for International Student Assessment (PISA) (2010a)

countries in both math and reading, was about the same as performance in the similar post-industrial countries in reading, and was worse than performance in the similar post-industrial countries in math.

We next disaggregate scores in the United States and in the six comparison countries by an approximation of the social class status of test takers, dividing them into six groups, from the least to the most advantaged. We refer to these as Group 1 (lowest social class), 2 (lower social class), 3 (lower-middle social class), 4 (upper-middle social class), 5 (higher social class), and 6 (highest social class). We also refer to Groups 1 and 2 together as disadvantaged students, to Groups 3 and 4 together as middle-class students, and to Groups 5 and 6 together as advantaged students.

There is no precise way to make social class comparisons between countries. PISA collects data on many characteristics that are arguably related to social class status, and also assembles them into an overall index. Although none of the possible indicators of social class differences is entirely satisfactory, we think one, the number of books in the home (BH), is probably superior for purposes of international test score comparisons, and we use it for our analysis. A very high fraction of students in both the PISA and TIMSS surveys answer the BH question, something less true for other important social class indicator questions asked on the student questionnaires. As we explain in greater detail below, we also examine whether other social class indicators, such as mother's education or PISA's overall index, in addition to BH, would produce meaningfully different results, and determine that they would not. We conclude that BH serves as a reasonable representation of social class (home) influences on students' academic performance.

Our examination of 2009 PISA scores, disaggregated by social class group, reveals that:

- In every country, students from more-advantaged social class groups outperform students from less-advantaged social class groups. The social class performance gap is large. In each country we study, the reading gap between the highest (Group 6) and the lowest (Group 1) social class groups is more than a full standard deviation. The math gap is also more than a full standard deviation in the United States and in four of the six comparison countries. In the other two, Canada and Finland, the gap is also large, almost a full standard deviation. The reading and math gaps are larger in France than in any country we studied.
- The reading and math gaps are smaller in the United States than in each of the three similar post-industrial countries we studied.
- The average U.S. scores in reading and math were about the same or lower than those in the six comparison countries in considerable part because a disproportionately greater share of U.S. students come from disadvantaged social class groups than do students in the six comparison countries.

TABLE 2A

Share of PISA 2009 sample in each social class group, by country

Social class group	Canada	Finland	Korea	France	Germany	U.K.	U.S.
Group 1 (Lowest)	9%	6%	5%	15%	12%	14%	20%
Group 2	13	11	9	17	13	16	18
Group 3	31	34	31	31	29	29	28
Group 4	21	23	23	18	19	18	16
Group 5	17	20	22	13	16	15	12
Group 6 (Highest)	9	6	9	7	10	8	6

Source: Authors' analysis of OECD Program for International Student Assessment (PISA) 2009 database for each country

TABLE 2B

Share of PISA 2009 sample in each social class group, for U.S., three top-scoring countries, and three similar post-industrial countries

Social class group	Average distribution for three top-scoring countries	Average distribution for three similar post-industrial countries	Distribution, U.S.
Group 1 (Lowest)	7%	14%	20%
Group 2	11	15	18
Group 3	32	30	28
Group 4	22	18	16
Group 5	20	15	12
Group 6 (Highest)	8	8	6

Source: Authors' analysis of OECD Program for International Student Assessment (PISA) 2009 database for each country

- If the United States had the same social class distribution as the average of the three top-scoring countries, or as the average of the three similar post-industrial countries, its average reading and math scores would have been higher than its reported averages.

Table 2A displays the share by social class group of the national samples for the United States and the six comparison countries.

Table 2B summarizes the data by grouping the comparison countries in Table 2A. Column (a) shows the average distribution by social class in the three top-scoring countries, and column (b) shows the average distribution by social class in the three similar post-industrial countries.

From these tables we can see that more U.S. 15-year-olds (37 percent[7]) are in the disadvantaged (Groups 1 and 2) social class groups than in any of the six comparison countries, and we can therefore see why comparisons that do not control for differences in social class distributions between countries may differ greatly from those that do. There are fewer U.S. students in the middle (Groups 3 and 4) social class groups than in the middle social class groups of the three similar post-industrial countries (Germany, France, and the United Kingdom), although the differences are small. Differences in the size of middle-class groups are larger when the United States is compared to the three top-scoring countries (Korea, Finland, and Canada). And in the advantaged (Groups 5 and 6) social class groups there are substantially fewer U.S. students

than there are in these groups in all six of the comparison countries.

Any meaningful comparison of average performance should be adjusted for these differences. To clarify why, consider two countries, in both of which affluent students score higher than poor students. Country A's most affluent (social class Group 6) students score higher than Country B's Group 6 students. Similarly, Country A's least advantaged (Group 1) students score higher than Country B's Group 1 students. Yet if the proportion of poor children in Country A is higher than the proportion of poor children in Country B, the average score of all students in Country A may be lower than the average score of all students in Country B, even though both affluent and poor students in Country A achieve at higher levels than socioeconomically similar students in Country B. Such apparent anomalies are termed "composition effects."

Before pursuing policies to address seemingly poor American student achievement in comparison to other nations, we should ask to what extent, if any, lower average U.S. performance is attributable to composition effects. In fact, a part, though small, of the apparently lower U.S. average performance is attributable to composition effects.

We can judge the importance of this composition effect by standardizing the social class distribution of the United States and the comparison countries. If we reweight the average country scores from Table 1, substituting the average social class weights of the top-scoring and similar post-industrial comparison countries from Table 2B, the country scores would be as shown in **Tables 3A-D**. **Tables 3A** and **3C** show what the 2009 PISA reading and math scores, respectively, would have been if each country had an identical social class distribution to that of the average of the top-scoring countries. **Tables 3B** and **3D** show what the 2009 PISA reading and math scores, respectively, would have been if each country had an identical social class distribution to that of the average of the similar post-industrial countries. **Figures A1** and **A2** (for reading) illustrate the data in Tables 3A and 3B; **Figures A3** and **A4** (for math) illustrate the data in Tables 3C and 3D.

The result of this reweighting is generally to increase scores in France and in the United States and to reduce scores in Korea. With reweighting, the U.S. average reading and math performance would still be below that of the top-scoring countries, although the U.S. deficit in reading in comparison to Canada would no longer be substantial. The U.S. average reading performance would now seem to be better than that in Germany or the United Kingdom, whereas before social class standardization the reading scores in these two countries were about the same as those in the United States.

Tables 3A and 3C show that if the U.S. PISA sample had the same social class weights as the average of the three top-scoring countries, and if the average performance of each social class group were the same as it was in actuality, the U.S. average reading score would not have been 500, but substantially better at 518, and the U.S. average math score would not have been 487, but better at 504.

Tables 3B and 3D show that if the U.S. PISA sample had the same social class weights as the average of the three similar post-industrial countries, and if the average performance of each social class group were the same as it was in actuality, the U.S. average reading score would not have been 500, but better at 509, and the U.S. average math score would not have been 487, but better at 495.

Tables 3A and 3B show that, in reading, if all countries in our study had the same social class composition as the average social class composition of the three top-scoring countries, or had the same social class composition as the average social class composition of the three similar post-industrial countries, the positive test score gap between the top-scoring countries and the United States would be cut in half, and the positive test score gap between the United States and similar post-industrial countries would at least double to become meaningful.

TABLE 3A

Overall average scale scores, reading, for U.S. and six comparison countries, PISA 2009 (with standardization for average social class distribution in top-scoring countries)

	TOP SCORING				SIMILAR POST-INDUSTRIAL				U.S.	U.S. VERSUS:	
	Canada	Finland	Korea	Average*	France	Germany	U.K.	Average*		Top-scoring average	Similar post-industrial average
National average reading score (from Table 1)	524	536	539	533	496	497	494	496	500	-33	4
National average reading score, standardized for top-scoring country average social class distribution	529	536	536	534	513	508	507	510	518	-16	9
Difference between social class standardized reading scores and actual average reading scores	5	0	-3	1	18	11	13	14	19		

* Simple (unweighted) average of three countries

Source: Authors' analysis of OECD Program for International Student Assessment (PISA) 2009 database for each country

Tables 3C and 3D show that, in math, if all countries in our study had the same social class composition as the average social class composition of the three top-scoring countries, or had the same social class composition as the average social class composition of the three similar post-industrial countries, the positive test score gap between the top-scoring countries and the United States would be cut by a third or more, and the positive test score gap between the similar post-industrial countries and the United States would also be cut by a third or more.

Tables 3A-D show how the U.S. average PISA reading and math scores might improve if the United States had the more favorable social class distributions of similar post-industrial countries. In Appendix A, we perform an opposite exercise, showing how much the scores of other countries might decline if they had the less favorable social class distribution of the United States. There is no single correct way to standardize scores by social class distribution. Other weighting methods generate somewhat different results, but the pattern is the same. Because of this distortion of average scores from social class composition, for the balance of this report, we focus on scores by social class group, not on average national scores.

Table 4 displays the 2009 reading and math scores for the United States and three similar post-industrial countries, disaggregated by comparable social class groups in each country.

In reading, in comparison to students in the three similar post-industrial countries, U.S. students from the lowest (Group 1) social class group scored substantially better than comparable social class students in each of the three similar post-industrial countries. U.S. students from the lower (Group 2) social class group performed better than

TABLE 3B

Overall average scale scores, reading, for U.S. and six comparison countries, PISA 2009 (with standardization for average social class distribution in similar post-industrial countries)

	TOP SCORING				SIMILAR POST-INDUSTRIAL				U.S.	U.S. VERSUS:	
	Canada	Finland	Korea	Average*	France	Germany	U.K.	Average*		Top-scoring average	Similar post-industrial average
National average reading score (from Table 1)	524	536	539	533	496	497	494	496	500	-33	4
National average reading score, standardized for similar post-industrial country average social class distribution	521	527	528	525	501	496	497	498	509	-17	11
Difference between social class standardized reading scores and actual average reading scores	-4	-8	-11	-8	5	-1	3	2	9		

* Simple (unweighted) average of three countries

Source: Authors' analysis of OECD Program for International Student Assessment (PISA) 2009 database for each country

comparable social class students in each of the three similar post-industrial countries. U.S. students in the lower-middle (Group 3) social class group performed better than comparable social class students in Germany and in the United Kingdom, and about the same as comparable social class students in France. U.S. students in the upper middle (Group 4) social class group performed about the same as comparable social class students in the three similar post-industrial countries. U.S. students in the higher (Group 5) social class group performed better than comparable social class students in Germany and in the United Kingdom, and about the same as comparable social class students in France. U.S. students in the highest (Group 6) social class group performed about the same as comparable social class students in the United Kingdom, better than comparable social class students in Germany, and worse than comparable social class students in France.

Tables 3A-B showed that the U.S. average reading score was higher than reported when social class distribution was controlled for. Table 4 shows that, in reading, U.S. students performed as well or better than students in the three similar post-industrial countries at every social class level. The only exception is students in France in the highest (Group 6) social class group, who performed better in reading than students in the United States.

In math, in comparison to students in the three similar post-industrial countries, U.S. students from the lowest (Group 1) social class group performed substantially better than comparable social class students in France and about the same as comparable social class students in Germany and the United Kingdom. U.S. students from the lower (Group 2) social class group performed about the same as comparable social class students in France and Germany and better than comparable social class students in the United Kingdom. In all other (Groups 3-6) social class groups, U.S. students performed sub-

TABLE 3C

Overall average scale scores, mathematics, for U.S. and six comparison countries, PISA 2009 (with standardization for average social class distribution in top-scoring countries)

	TOP SCORING				SIMILAR POST-INDUSTRIAL				U.S.	U.S. VERSUS:	
	Canada	Finland	Korea	Average*	France	Germany	U.K.	Average*		Top-scoring average	Similar post-industrial average
National average math score (from Table 1)	527	541	546	538	497	513	492	501	487	-50	-13
National average math score, standardized for top-scoring country average social class distribution	531	541	543	538	513	522	504	513	504	-34	-9
Difference between social class standardized math scores and actual average reading scores	4	0	-3	0	17	10	11	13	17		

* Simple (unweighted) average of three countries

Source: Authors' analysis of OECD Program for International Student Assessment (PISA) 2009 database for each country

stantially worse than comparable social class students in Germany, and about the same as comparable social class students in the United Kingdom. U.S. students in the upper-middle (Group 4) and highest (Group 6) social class groups performed substantially worse than comparable social class students in France, and U.S. students in the higher (Group 5) social class group performed worse than comparable social class students in France.

Unlike in reading, however, in math U.S. students underperformed students from middle and advantaged (Groups 3-6) social class groups in France and Germany, and mostly performed about the same as students from similar social class groups in the United Kingdom. Only in a comparison with the lowest (Group 1) social class students in France were comparable social class U.S. students substantially superior in math performance.

Table 4 also displays the test score gradient (commonly referred to as the "achievement gap"), measured in two ways: the gap in average scores between students in Group 1 and students in Group 6, and the gap in average scores between students in Group 2 and students in Group 5.

In reading, the Group 1/Group 6 achievement gap is smaller in the United States than in the three similar post-industrial countries, and much smaller than in France. The Group 2/Group 5 reading achievement gap is smaller in the United States than in France or the United Kingdom.[8] In math, the Group 1/Group 6 achievement gap is smaller in the United States than in France or Germany, and about the same as in the United Kingdom. The Group 2/Group 5 math achievement gap is smaller in the United States than in each of the similar post-industrial countries.

TABLE 3D

Overall average scale scores, mathematics, for U.S. and six comparison countries, PISA 2009 (with standardization for average social class distribution in similar post-industrial countries)

	TOP SCORING				SIMILAR POST-INDUSTRIAL				U.S.	U.S. VERSUS:	
	Canada	Finland	Korea	Average*	France	Germany	U.K.	Average*		Top-scoring average	Similar post-industrial average
National average math score (from Table 1)	527	541	546	538	497	513	492	501	487	-50	-13
National average math score, standardized for similar post-industrial country average social class distribution	523	534	533	530	502	511	495	502	495	-35	-7
Difference between social class standardized math scores and actual average reading scores	-3	-6	-13	-8	5	-2	2	2	8		

* Simple (unweighted) average of three countries

Source: Authors' analysis of OECD Program for International Student Assessment (PISA) 2009 database for each country

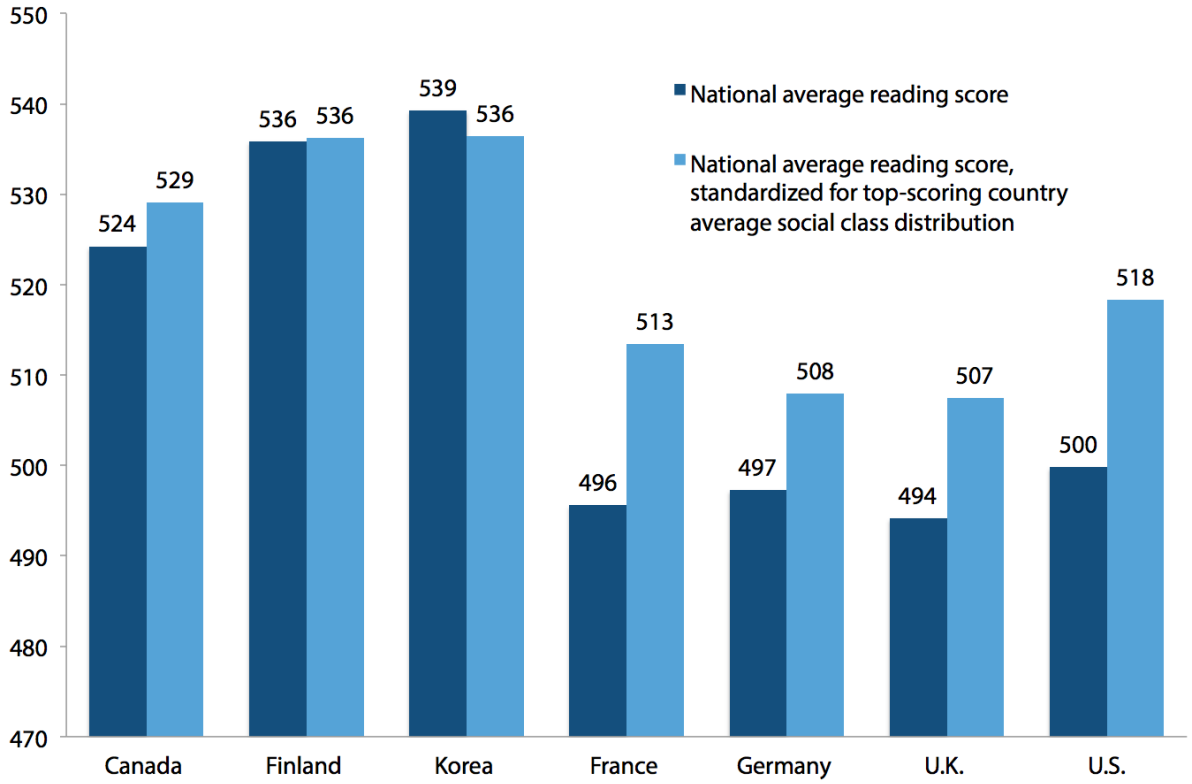
Careful examination of these gradients, however, should serve as a warning to be cautious about interpretation of "achievement gaps," the subject of frequent policy comment in the United States. One interpretation of these gradients, mostly larger in the similar post-industrial countries than in the United States, suggests that social class has a bigger impact on reading and math performance in the similar post-industrial countries than it does in the United States. Perhaps this is because the United States has a more equal school system than have the similar post-industrial countries, or because non-school social class characteristics have a bigger impact in the similar post-industrial countries than they do in the United States. Either of these explanations is at variance with commonplace assumptions in U.S. policy discussion. This finding is especially noteworthy because income inequality is probably larger in the United States than in the similar post-industrial countries.

However, having a more equal school system is not necessarily the same as having a superior school system. Consider the Group 2/Group 5 gradients for the United States and France: In reading, the U.S. gap is smaller than the gap in France. This is attributable to the United States having higher reading achievement in Group 2 and about the same reading achievement in Group 5. This seems to be a desirable relative (to France) outcome for the United States. But in math, the smaller U.S. gap is attributable to Group 2 mathematics achievement that is about the same in the two countries, with Group 5 mathematics achievement that is lower in the United States than in France. Generating a smaller gap by having lower achievement in the higher social class group is probably not a result most policymakers would seek.

The U.S.-Germany reading gradient comparison is even more favorable to the United States than the U.S.-France gradient comparison, with U.S. achievement higher both for Group 2 and Group 5 students. Because the Group 2

FIGURE A1

Average national reading scores, actual and re-weighted using top-scoring country average social class group distribution, for U.S. and six comparison countries, PISA 2009



Source: Authors' analysis of OECD Program for International Student Assessment (PISA) 2009 database for each country

U.S. superiority is greater than the Group 5 superiority, the U.S. gap is smaller. This is a desirable result.

But in math, the smaller U.S. gap relative to the German gap is attributable to Group 2 scores that are about the same in the two countries while Group 5 scores are substantially lower in the United States than in Germany. Although the United States has a smaller achievement gap, this is not a desirable result.

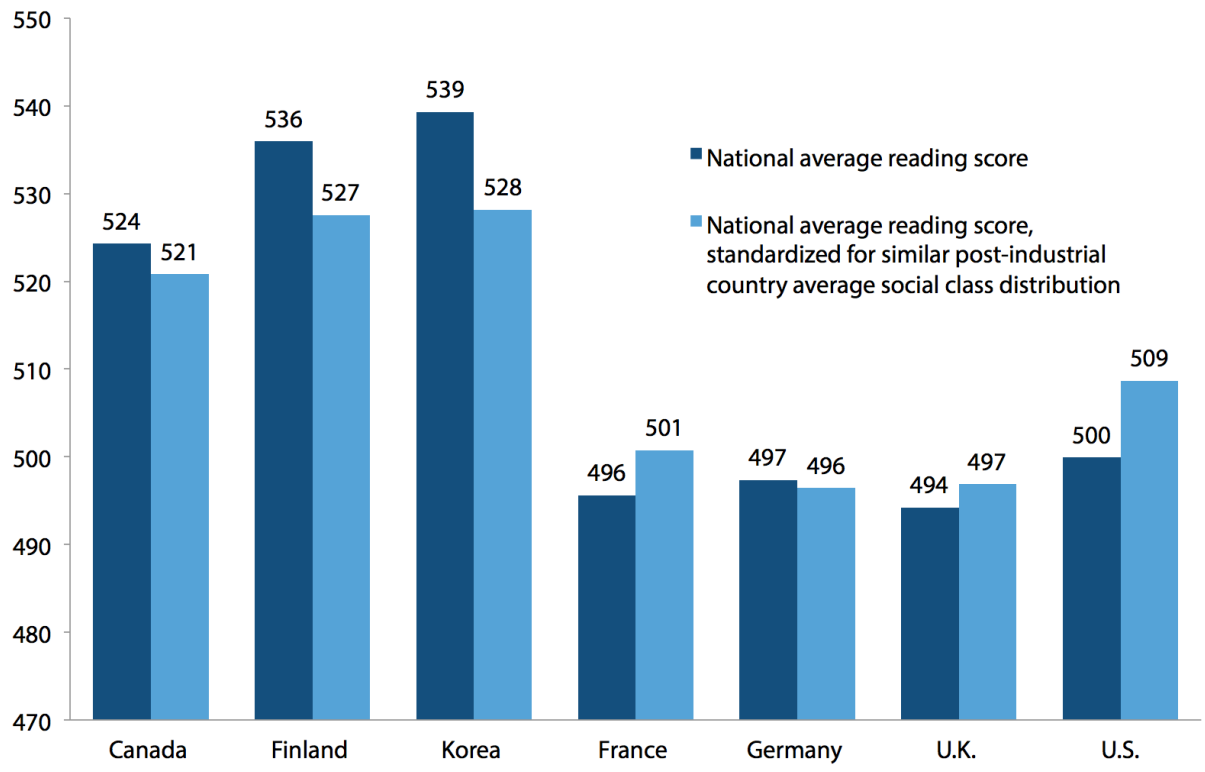
Comparing the U.S. and U.K. gradients, in reading the result is similar to that in the German comparison—desirable for the United States, because U.S. Group 2 achievement is higher than that in the United Kingdom, while U.S. Group 5 achievement is also higher than in the United Kingdom, but not as much so. In math, U.S. achievement in Group 2 is higher than that in the United Kingdom, while Group 5 achievement in the two countries is about the same. This, too, is a desirable result for the United States, but not as desirable as it would be if Group 5 achievement were higher as well.

Table 5 displays the 2009 reading and math scores for the United States and three top-scoring countries, disaggregated by comparable social class groups in each country.

In reading, disadvantaged (Groups 1 and 2) students in the U.S. score substantially worse than comparable students in the three top-scoring countries, the only exception being the lowest (Group 1) social class students, where U.S. students score worse but not substantially worse than their social class counterparts in Canada. Likewise for middle (Groups 3 and 4) social class students: U.S. students score worse than comparable students in Canada and substantially worse than comparable students in Finland and Korea. Higher (Group 5) social class stu-

FIGURE A2

Average national reading scores, actual and re-weighted using similar post-industrial country average social class group distribution, for U.S. and six comparison countries, PISA 2009



Source: Authors' analysis of OECD Program for International Student Assessment (PISA) 2009 database for each country

dents in the United States score about the same as comparable social class students in the three top-scoring countries, while the highest (Group 6) social class students in the United States score worse than comparable social class students in Finland and Korea and about the same as comparable social class students in Canada.

In comparing the United States and the three top-scoring countries in math, the picture is consistent across all social class groups and countries: U.S. students score substantially worse than comparable students in each social class group in the three top-scoring countries, the exception being that U.S. higher social class (Group 5) students score worse than comparable social class students in Canada.

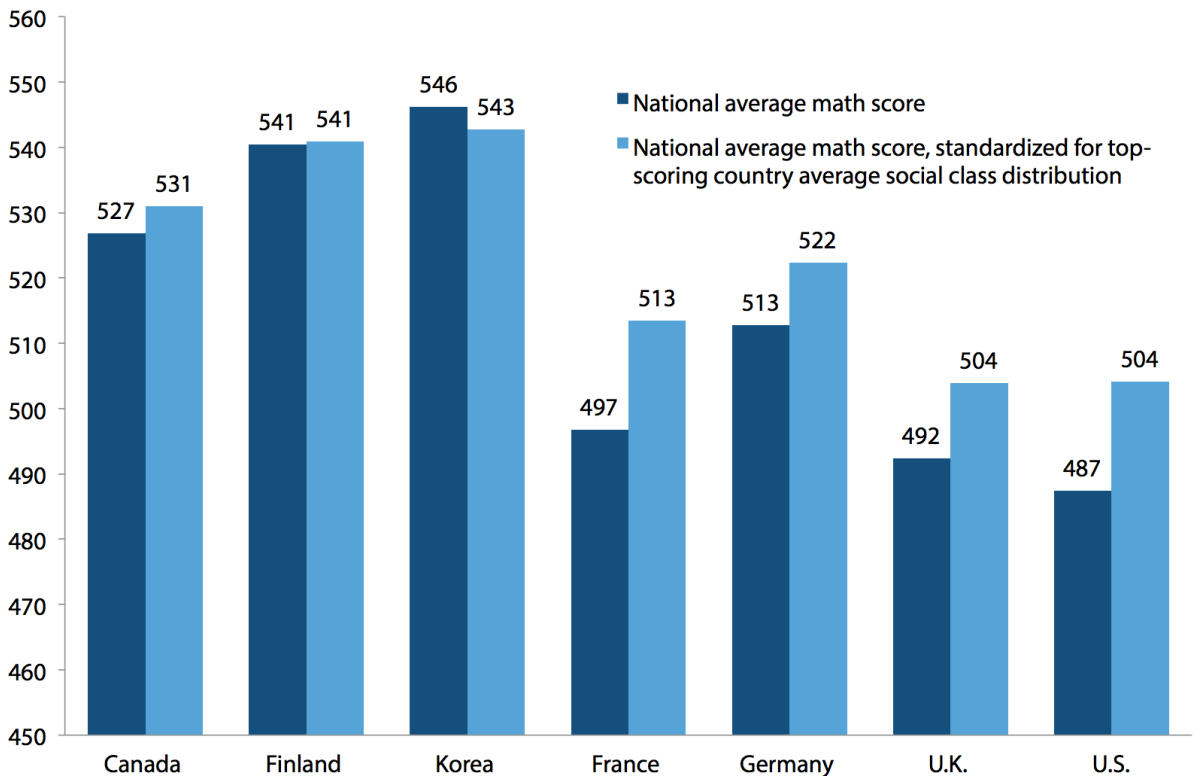
Table 5 also displays the test score gradients between advantaged and disadvantaged students in the United States and the top-scoring countries.

Unlike the gradients in the similar post-industrial countries, the gradients in the top-scoring countries are generally smaller than those in the United States. In reading, the Group 6/Group 1 gap is smaller in Canada and in Finland than in the United States and about the same in Korea as in the United States. The Group 5/Group 2 reading gradient is smaller in Finland than in the United States and much smaller in Canada and Korea than in the United States.

In math, the Group 6/Group 1 gradient is much smaller in Canada and Finland than in the United States, as is the Group 5/Group 2 math gradient in Finland. The Group

FIGURE A3

Average national math scores, actual and re-weighted using top-scoring country average social class group distribution, for U.S. and six comparison countries, PISA 2009



Source: Authors' analysis of OECD Program for International Student Assessment (PISA) 2009 database for each country

5/Group 2 math gradients in Canada and Korea are smaller than in the United States.

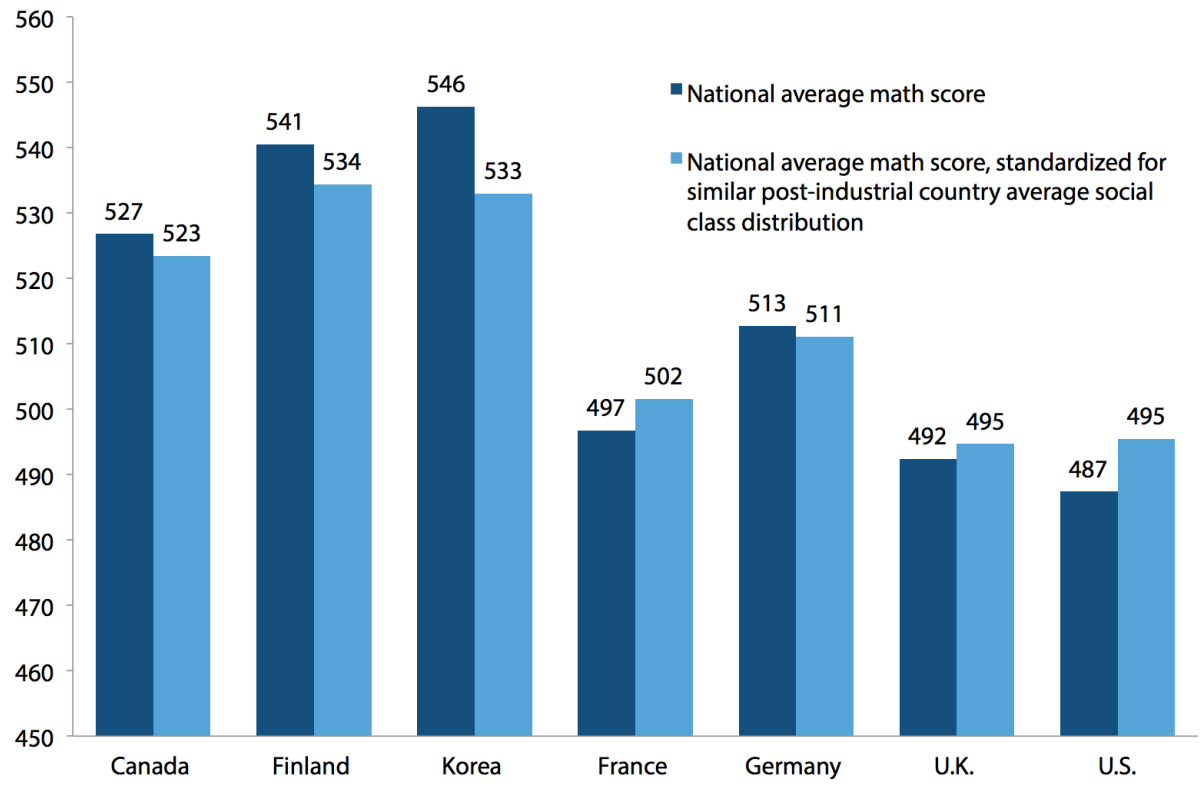
What stands out most, however, is the unusually large gap in achievement between Korean students in Group 6 and those in Group 1. This gradient of 149 scale points is larger than in any other comparison we have made, and results both from the unusually low relative performance in math of Korean students in Group 1 and unusually high relative performance in math of Korean students in Group 6. Although the lowest (Group 1) social class students in Korea score substantially better than similar social class students in the United States, the relative advantage of Korean performance is much more pronounced at the highest (Group 6) social class level.

We cannot say whether this Korea–United States difference is attributable to the United States having a more equal school system than does Korea, or because non-school characteristics of the highest social class students have a bigger positive impact on students in Korea than on students in the United States. For example, widely reported access to out-of-school tutoring may have an unusually large impact on the highest social class students in Korea.

The comparisons described in this part of the report show that, to some extent, the widely reported disparity between the performance of U.S. students and that of comparable countries' students on the PISA is attributable to the U.S. sample of test takers being more heavily weighted toward disadvantaged students than the samples of comparable countries. Although adjustment for these social class differences does not eliminate the gap between the performance of United States and top-scoring country students, it narrows the gap. And relative to the perform-

FIGURE A4

Average national math scores, actual and re-weighted using similar post-industrial country average social class group distribution, for U.S. and six comparison countries, PISA 2009



Source: Authors' analysis of OECD Program for International Student Assessment (PISA) 2009 database for each country

ance of students in similar post-industrial countries, the performance of U.S. students in many cases no longer seems deficient once social class composition is taken into account.

In this connection, we note here but reserve for detailed discussion in Part IV an apparent flaw in the 2009 U.S. PISA sampling methodology. Although the U.S. sample included disadvantaged students in appropriate proportion to their actual representation in the U.S. 15-year-old population, the U.S. sample included a disproportionate number of disadvantaged students who were enrolled in schools with unusually large concentrations of such students. Because, after controlling for student social class status, students from families with low social class status will perform more poorly in schools with large concentrations of such students, this sampling flaw probably reduced the reported average score of students in the bottom social class groups (perhaps Groups 1-3). However, with available data, we cannot say to what extent this occurred. We do conclude, however, that this distortion probably depressed the reported average scores of U.S. students beyond the composition effect discussed in this section, artificially reducing the reported U.S. average score and its international ranking.

A consistent pattern in the 2009 PISA scores is the better performance of U.S. students on the reading than on the math test, relative to the comparison countries. **Table 6** displays this pattern.

For each social class group in each comparison country, the table shows the difference between the reading gap for a U.S. comparison and the math gap for a U.S. comparison. For example, for the lowest (Group 1) social class, the Canada–U.S. reading gap is 17 scale points

TABLE 6

Reading vs. math, U.S. compared with other countries, PISA 2009

	U.S. VERSUS:					
	Canada	Finland	Korea	France	Germany	U.K.
Group 1 (Lowest)	19	32	-1	18	28	19
Group 2	8	18	9	9	18	7
Group 3	17	19	15	14	27	11
Group 4	19	18	26	14	32	15
Group 5	14	15	30	18	32	7
Group 6 (Highest)	15	22	36	11	35	4

Note: Numbers in this table are the reading gap less the math gap for each social class group. The reading (math) gap is the U.S. average reading (math) score for a given social class group less the comparison country's reading (math) score for that social class group.

Source: Authors' analysis of OECD Program for International Student Assessment (PISA) 2009 database for each country

TABLE 4

Scale scores by social class group for U.S. and similar post-industrial countries, PISA 2009

	France	Germany	U.K.	U.S.
Reading				
Group 1 (Lowest)	403	413	424	442
Group 2	458	455	455	471
Group 3	498	496	490	504
Group 4	533	523	522	529
Group 5	559	555	555	563
Group 6 (Highest)	573	551	562	563
Gap (Group 6 – Group 1)	170	137	138	121
Gap (Group 5 – Group 2)	101	100	100	93
Math				
Group 1 (Lowest)	413	433	435	434
Group 2	460	466	455	464
Group 3	498	509	487	491
Group 4	529	535	517	510
Group 5	562	571	547	548
Group 6 (Highest)	569	570	551	548
Gap (Group 6 – Group 1)	156	137	116	114
Gap (Group 5 – Group 2)	102	104	92	84

Source: Authors' analysis of OECD Program for International Student Assessment (PISA) 2009 database for each country

TABLE 5

Scale scores by social class group for U.S. and top-scoring countries, PISA 2009

	Canada	Finland	Korea	U.S.
Reading				
Group 1 (Lowest)	459	466	461	442
Group 2	492	495	501	471
Group 3	518	523	529	504
Group 4	543	552	546	529
Group 5	561	571	564	563
Group 6 (Highest)	567	572	581	563
Gap (Group 6 – Group 1)	108	106	119	121
Gap (Group 5 – Group 2)	70	75	63	93
Math				
Group 1 (Lowest)	471	490	452	434
Group 2	493	507	504	464
Group 3	521	528	531	491
Group 4	543	552	553	510
Group 5	560	570	579	548
Group 6 (Highest)	567	580	602	548
Gap (Group 6 – Group 1)	96	90	149	114
Gap (Group 5 – Group 2)	67	63	75	84

Source: Authors' analysis of OECD Program for International Student Assessment (PISA) 2009 database for each country

(from Table 5, the U.S. Group 1 reading score is 442 and the Canadian Group 1 reading score is 459). The Canada–U.S. math gap is 37 scale points (from Table 5, the U.S. Group 1 math score is 434 and the Canadian Group 1 math score is 471). The difference between the reading gap of 17 scale points and the math gap of 37

scale points is the 19 scale points shown in Table 6 for Group 1, Canada. Wherever a positive number appears in Table 6, the reading gap is smaller than the math gap. Note that a positive number does not signify that U.S. students perform better in reading than students in the same social class group in a comparison country, or better in reading but not in math; it may mean that, or it may mean that the U.S. comparative deficit is less in reading than in math for that particular social class group and country because the reading deficit is smaller than the math deficit.

Table 6 shows that, on average, and for almost every social class group U.S. students do relatively better in reading than in math, compared to students in both the top-scoring and the similar post-industrial countries. The only exceptions to this pattern are with respect to social class Group 1 in Korea and to social class Groups 2, 5, and 6 in the United Kingdom. In these four cases, the reading and math gaps are about the same. In all other comparisons (for each social class group in each of the six comparison countries), the United States does relatively better in reading than in math, either because the U.S. reading score is higher than the reading score for the same social class group in a comparison country and the U.S. math score is less higher or lower, or because the U.S. reading score is lower than the reading score in the same social class group in a comparison country by a lesser amount than the U.S. math score is lower.

Part III. PISA trends from 2000 to 2009

Data are now available for four administrations of PISA – 2000, 2003, 2006 and 2009. Score trends over this decade may seem surprising. We would ordinarily expect instruction to be more difficult when the concentration of disadvantaged students increases. Yet while the social class composition of the national PISA sample deteriorated more in the United States than in any other country, disadvantaged U.S. students nonetheless saw their scores improve, while scores of similarly disadvantaged students in countries to which the United States is frequently compared have been declining. PISA reported that the U.S. average reading score was about the same in 2009 as it had been in 2000, but if U.S. social class composition had not deteriorated, the average U.S. reading score would have improved from 2000 to 2009. PISA reported that the U.S. average math score was worse in 2009 than in 2000, but this was all because of deteriorating social class composition. If this deterioration had not occurred, U.S. average math performance would have been about the same in 2009 as it had been in 2000.

The test score gaps between disadvantaged students in the United States and in top-scoring countries generally narrowed, but the gaps between advantaged students in the United States and in these top-scoring countries widened in some cases. In comparison to similar post-industrial countries, the United States also narrowed the gap more at the bottom than at the top, and in some cases ended the decade with clear superiority over similar social class groups toward the bottom of the scale. This section explores these findings in greater detail.

Score trends over time are as important for policy purposes as score levels at the current time. We want to know not only in which countries adolescents perform better than in other countries, but also whether there are socioeconomic factors or educational policies and practices that are causing a country's performance to improve or deteriorate. If one country has lower 2009 PISA scores than another, but if scores in the lower-scoring country have been improving over the previous decade while scores in the higher-scoring country have been declining, policymakers in the lower-scoring country might be ill-advised to look exclusively to the higher-scoring country for model school improvement policies. At the very least, policymakers should attempt to understand why the higher-scoring country's superior achievement appears, at least to some extent, to be unsustainable.

PISA has been administered every three years since 2000, and the multiple years of data provide policymakers an

opportunity to make more useful judgments than would be allowed by a single year of data. Unfortunately, there are no U.S. reading data for 2006 because of an error in test administration.[9] Thus, we can look at changes in U.S. students' math performance on PISA from 2000 to 2003, to 2006, and to 2009, but at reading performance only from 2000 to 2003 and then to 2009.

Students who were 15 years old and took the PISA in 2000 would have been affected by their families' social, economic, and community environments beginning in about 1985, and would have entered school in about 1990. PISA score changes from 2000 to 2003 could have been influenced by socioeconomic or instructional or other educational changes that took place anywhere from the mid-1980s to 2003. Likewise, PISA score changes from 2003 to 2006 could have been influenced by socioeconomic or instructional changes that took place anywhere from the late 1980s to 2006. And PISA score changes from 2006 to 2009 could have been influenced by socioeconomic or instructional changes that took place anywhere from the mid-1990s to 2009.

In this report, we are unable to attribute causes to trends in scores; we can only describe them. We review trends in reading and math for the United States and each of the six comparison countries in the discussion and tables that follow.

As was the case when we examined comparative score levels in 2009, our main conclusion from this review is that there are few consistent patterns in these score trends that can be used to inspire policy. Simplistic judgments based on selective or overly generalized data can (and do) mask critical aspects of U.S. relative performance, and they can support policy changes that can undermine U.S. sources of strength and exacerbate U.S. sources of weakness.

As in the previous section of this report, we focus on trends by social class group, because changes over time in the composition of a country's test takers by social class can affect a country's average score while masking real changes (or lack of change) in the performance of that country's students. Composition effects can distort changes over time as well as comparisons between countries at a given time.

In fact, the proportion of students sampled in different social class groups from 2000 to 2009 in the United States and in the six comparison countries has changed, and these shifts influence changes in the overall average score of each country over time.

It is made somewhat more difficult to understand these changes because PISA modified its books-in-the-home (BH) group definitions after the 2000 assessment. **Table 7** displays these changed definitions.[10]

TABLE 7

PISA group definitions by books in the home

	NUMBER OF BOOKS IN HOME	
	2000	**2003 and after**
Group 0	0	–
Group 1	1–10	0–10
Group 2	11–50	11–25
Group 3	51–100	26–100
Group 4	101–250	101–200
Group 5	251–500	201–500
Group 6	>500	>500

Source: OECD Program for International Student Assessment (PISA) 2000, 2003, 2006, and 2009 databases

We can make some comparisons of social class distributions of test takers in 2000 and 2009 because four categories are consistent over this period: a combination of Groups 0 and 1, which includes test takers from homes with 10 books or fewer; a combination of Groups 2 and 3, which includes test takers from homes with 11 to 100 books; a combination of Groups 4 and 5, which includes test takers from homes with 101 to 500 books; and Group 6, which includes test takers from homes with more than 500 books.

TABLE 8A

Changes in PISA sample social class composition by books in the home, U.S. and six comparison countries, 2000–2009 (percentage points)

	Canada	Finland	Korea	France	Germany	U.K.	U.S.
0–10 books	+2	-2	-3	+3	+4	+5	+7
11–100 books	+6	-4	-1	+3	0	+3	+5
101–500 books	-5	+6	+2	-5	-2	-3	-7
>500 books	-4	0	+2	-1	-2	-5	-4

Source: Authors' analysis of OECD Program for International Student Assessment (PISA) 2000 and 2009 databases for each country

TABLE 8B

Changes in PISA sample social class composition by books-in-the-home group, U.S. and six comparison countries, 2003–2009 (percentage points)

	Canada	Finland	Korea	France	Germany	U.K.	U.S.
Group 1 (Lowest)	+2	+1	-1	+6	+5	+5	+7
Group 2	+2	-2	-2	+1	0	+2	+2
Group 3	+1	-3	-2	-3	-1	-1	-3
Group 4	-1	+1	-2	-2	-2	0	-3
Group 5	-1	+4	+4	0	-1	-3	-2
Group 6 (Highest)	-3	0	+3	-1	-2	-2	-2
Disadvantaged (Groups 1 and 2)	+4	-1	-3	+6	+5	+7	+10
Middle class (Groups 3 and 4)	0	-2	-4	-5	-3	-1	-6
Advantaged (Groups 5 and 6)	-4	+3	+6	-1	-3	-5	-4

Source: Authors' analysis of OECD Program for International Student Assessment (PISA) 2000 and 2009 databases for each country, with authors' interpolations for 2000 social class composition to match 2009 books-in-the-home groupings

Table 8A shows how the distribution of test takers by these four books-in-the-home categories in each of the seven countries changed from 2000 to 2009.

The table shows that the share of students whose homes had the fewest (0-10) books declined in Finland and Korea, but increased in Canada, France, Germany, the United Kingdom, and, most of all, the United States. The share of students from homes with only 11-100 books also increased in the United States and in Canada as well. Correspondingly, the share of students whose homes had more than 100 books increased in Finland and Korea, but declined everywhere else, with the largest decline in the United States. By these measures of change in the sample proportions of students from homes with fewer and more books, U.S. students' average social class declined more than the average social class of any of the comparison countries from 2000 to 2009, with the United Kingdom a close second. Finland and Korea's average social class increased.

Because the BH categories remained consistent from 2003 onward, **Table 8B** shows how the distribution of test takers by social class in these countries changed from 2003 to 2009.

We can see from Table 8B that, during the six-year period 2003–2009, the average social class of the test-taking samples in Canada, in the three similar post-industrial countries (France, Germany, and the United Kingdom),

TABLE 9A

Reading score changes, scale scores by social class group for U.S. and similar post-industrial countries, PISA 2000–2009

	FRANCE			GERMANY			U.K.			U.S.		
	2000	2009	Change	2000	2009	Change	2000	2009	Change	2000	2009	Change
Group 1 (Lowest)	430	403	-27	361	413	52	440	424	-17	418	442	23
Group 2	464	458	-7	404	455	52	470	455	-16	455	471	15
Group 3	503	498	-5	465	496	31	508	490	-19	499	504	5
Group 4	526	533	8	502	523	21	539	522	-17	528	529	1
Group 5	553	559	6	536	555	19	565	555	-10	556	563	7
Group 6 (Highest)	548	573	26	549	551	1	577	562	-15	560	563	3
National average reading score	505	496	-9	484	497	13	523	494	-29	504	500	-5

Source: Authors' analysis of OECD Program for International Student Assessment (PISA) 2000 database, with authors' interpolations of average test scores; Tables 1 and 4 for 2009 data

and in the United States declined, with the U.S. decline larger than in any of the comparison countries.

Because of such social class compositional changes, comparisons of test score trends over time by social class group provide more useful information to policymakers than comparisons of total average test scores at one point in time or even of changes in total average test scores over time.

For reading and math, we examine trends in the United States by BH categories compared to the six comparison countries for the 2000 to 2009 period. The paths by which performance changed from 2000 to 2009 varied by country, so an investigation of why these 2000 to 2009 changes occurred in specific countries should also examine disaggregated scores. For the United States, because no data are available for reading in 2006, such an investigation should disaggregate the reading trends by examining the 2000 to 2003 and 2003 to 2009 periods separately. For mathematics, a similar investigation would be appropriate, with the addition of disaggregating trends for the 2003 to 2006 and 2006 to 2009 periods.

In the next series of tables, we show how, for each social class group, PISA achievement in reading and math changed in the United States and in each of the comparison countries from 2000 to 2009. Because, as noted above, PISA changed its books-in-the-home categories in 2003, social class groups in 2000 do not exactly match the categories in 2009. Thus, to make an estimate of average social class group score changes from 2000 to 2009, we interpolate average scores for books-in-the-home categories in 2000 in order to create average test scores by social class groups that are comparable to those in 2009.[11] We use these estimates to calculate test score differences by social class groups from 2000 to 2009.

Reading, 2000–2009

Table 9A displays how reading achievement changed from 2000 to 2009 in the United States and the three similar post-industrial countries.

Table 9B displays data on how reading gaps between U.S. students and comparable social class students in the three similar post-industrial countries changed from 2000 to 2009. (Positive numbers describe gains for U.S. performance relative to the performance of comparison countries. Negative numbers describe deteriorated U.S. performance relative to that of comparison countries.)

TABLE 9B

Reading score gap changes, U.S. vs. similar post-industrial countries, PISA 2000–2009

	GAP CHANGES, U.S. VERSUS:		
	France	Germany	U.K.
Group 1 (Lowest)	+50	-29	+40
Group 2	+22	-36	+31
Group 3	+10	-26	+24
Group 4	-7	-19	+18
Group 5	+1	-12	+17
Group 6 (Highest)	-23	+1	+18

Note: Numbers in this table take the 2009 U.S. average score for a social class group, less the 2009 comparison country's average score for the same social class group, and subtract from this result the 2000 U.S. average score for that social class group, less the 2000 comparison country's average score for the same social class group.

Source: Authors' analysis of OECD Program for International Student Assessment (PISA) 2000 database, with authors' interpolations of average test scores; Tables 1 and 4 for 2009 data

Considering the full 2000 to 2009 period, U.S. reading scores improved for disadvantaged social class (Groups 1-2) students, including a substantial improvement for the lowest social class (Group 1); U.S. reading scores were about the same for middle-class and advantaged social class (Groups 3-6) students.

Considering trends in the three similar post-industrial countries in the full 2000 to 2009 period:

- In France, reading scores declined substantially for the lowest social class (Group 1) students, improved for upper-middle social class (Group 4) students, improved substantially for the highest social class (Group 6) students, and were mostly unchanged for lower-middle and higher social class (Groups 3 and 5) students. Thus, whereas in 2000 U.S. disadvantaged social class (Groups 1-2) students performed below comparable French students, in 2009 these students in the United States performed better than disadvantaged students in France and, in the case of the lowest social class (Group 1) students, substantially better. Whereas in 2000 the highest social class (Group 6) students in the United States performed better than comparable French students, in 2009 they performed worse. Middle and higher social class students (Groups 3-5) in the United States and France performed at about the same level in both years.[12]
- In Germany, reading scores were mostly unchanged from 2000 to 2009 for the highest social class (Group 6) students but improved substantially for other social class (Groups 1-5) students. There were extraordinarily large gains—half a standard deviation—for disadvantaged social class group (Groups 1-2) students. Thus, although U.S. students still had higher reading scores than German students in each social class group in 2009 (except for upper-middle social class [Group 4] students, who scored about the same in the two countries in 2009), and although the lowest social class (Group 1) students in the United States continued to perform substantially better than comparable German students, German students closed the gap in all social class groups (except for Group 6) from 2000 to 2009.
- In the United Kingdom, reading scores declined in every social class group, with substantial declines for lower-middle social class (Group 3) students. Thus, whereas in 2000 U.S. students performed worse than U.K. students in each social class group, by 2009 the lowest social class (Group 1) students in the United States performed substantially better than comparable students in the United Kingdom, and lower, lower-middle, and higher social class (Groups 2, 3, and 5) students in the United States performed better than comparable social class students in the United Kingdom. Upper-middle and the highest social class (Groups 4 and 6) students in the United States performed about the same in 2009 as comparable social class students in the United Kingdom.

Table 10A displays how reading achievement changed from 2000 to 2009 in the United States and the three top-scoring countries.

TABLE 10A

Reading score changes, scale scores by social class group for U.S. and top-scoring countries, PISA 2000–2009

	CANADA			FINLAND			KOREA			U.S.		
	2000	2009	Change	2000	2009	Change	2000	2009	Change	2000	2009	Change
Group 1 (Lowest)	467	459	-8	497	466	-31	464	461	-3	418	442	23
Group 2	490	492	1	514	495	-19	490	501	11	455	471	15
Group 3	522	518	-5	534	523	-11	518	529	11	499	504	5
Group 4	542	543	1	558	552	-6	532	546	14	528	529	1
Group 5	560	561	1	575	571	-4	546	564	19	556	563	7
Group 6 (Highest)	563	567	4	581	572	-9	556	581	25	560	563	3
National average reading score	534	524	-10	546	536	-11	525	539	15	504	500	-5

Source: Authors' analysis of OECD Program for International Student Assessment (PISA), 2000 database, with authors' interpolations of average test scores; Tables 1 and 5 for 2009 data

Table 10B displays the data on how reading gaps between U.S. students and comparable social class students in the top-scoring countries changed from 2000 to 2009. (Positive numbers describe gains for U.S. performance relative to the performance of comparison countries. Negative numbers describe deteriorated U.S. performance relative to that of comparison countries.)

TABLE 10B

Reading score gap changes, U.S. vs. top-scoring countries, PISA 2000–2009

	GAP CHANGES, U.S. VERSUS:		
	Canada	Finland	Korea
Group 1 (Lowest)	+31	+54	+26
Group 2	+14	+34	+4
Group 3	+10	+16	-6
Group 4	+0	+7	-13
Group 5	+6	+12	-11
Group 6 (Highest)	-1	+11	-22

Note: Numbers in this table take the 2009 U.S. average score for a social class group, less the 2009 comparison country's average score for the same social class group, and subtract from this result the 2000 U.S. average score for that social class group, less the 2000 comparison country's average score for the same social class group.

Source: Table 10A

Considering trends in the three top-scoring countries in the full 2000 to 2009 period:

- In Canada, reading scores declined for the lowest social class (Group 1) students, and were mostly unchanged for all others (Groups 2-6). Thus, while the lowest social class (Group 1) students in the United States still performed below comparable social class students in Canada, the gap between these U.S. and Canadian students was cut by two-thirds during this period. Gaps were also narrowed for lower- and lower-middle-class (Groups 2 and 3) students, while for upper-middle and advantaged social class (Groups 4-6) students, the gap was mostly unchanged from 2000 to 2009.
- In Finland, reading scores declined for disadvantaged, lower-middle, and the highest social class (Groups 1-3 and 6) students, with substantial declines for disadvantaged social class (Groups 1-2) students. Reading scores for upper-middle and higher social class (Group 4 and 5) students were about the same in both years. U.S. disadvantaged and middle social class (Groups 1-4) students still scored substantially below comparable students in Finland in 2009. The highest social class (Group 6) students also scored below com-

parable students in Finland, but higher social class (Group 5) students now scored about the same in the United States and Finland. The U.S.-Finland reading gap was cut by about two-thirds for disadvantaged social class (Groups 1 and 2) students, was cut in half for lower-middle and advantaged social class (Groups 3, 5, and 6) students, and by about a third for upper-middle social class (Group 4) students from 2000 to 2009.

- In Korea, reading scores improved for lower and middle social class (Groups 2-4) students and improved substantially for advantaged social class (Groups 5-6) students. Korean reading scores remained the same for lowest social class (Group 1) students. U.S. lowest social class (Group 1) students narrowed substantially (but did not eliminate) their negative performance gap relative to comparable students in Korea, but the U.S. negative performance gap grew for upper-middle and advantaged social class (Groups 4-6) students, with a substantial growth in this gap for the highest social class (Group 6) students. While U.S. higher social class (Group 5) students outperformed comparable social class students in Korea in 2000, by 2009 this social class group performed about the same in the two countries.

Thus, although U.S. students still scored below each of the three top-scoring countries in reading in almost all social class groups, U.S. students narrowed the gap in many groups from 2000 to 2009. Of particular note is the substantial gap closing between the lowest social class (Group 1) students in the United States and in each top-scoring country.

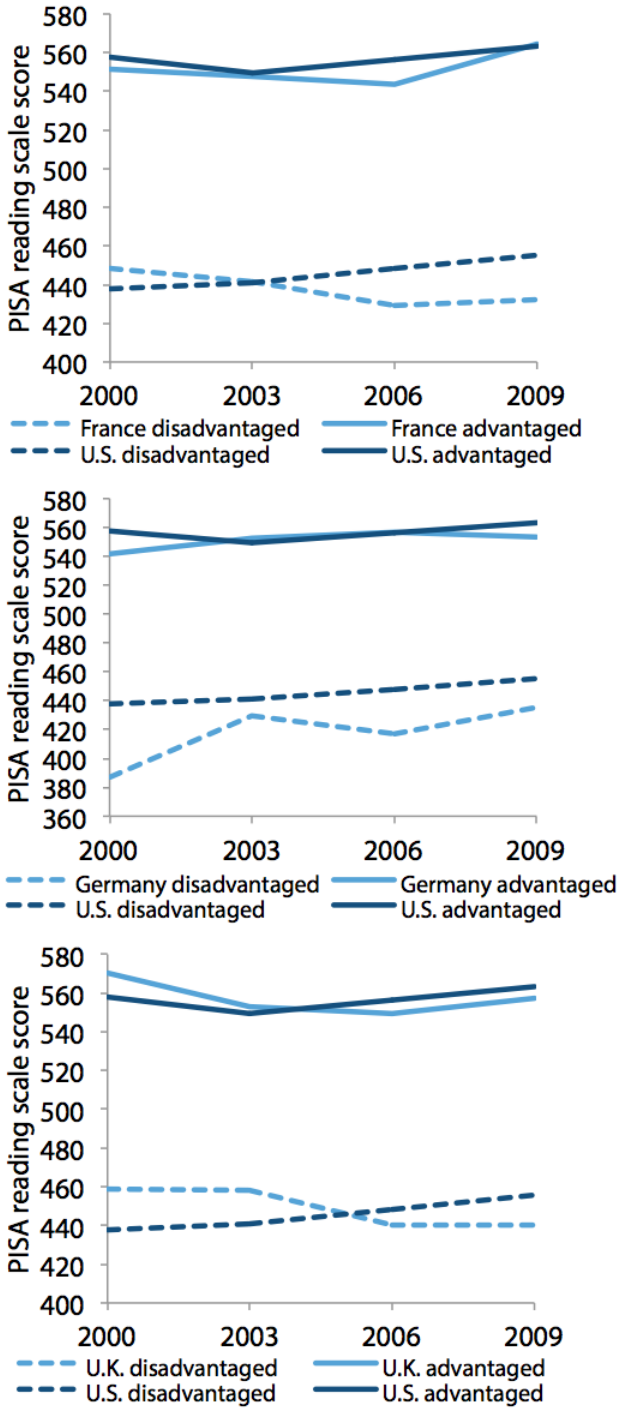
The 2000 to 2009 trends just described are not always (indeed, rarely) linear. For each social class group in each country, performance may have risen and then fallen during the period, making an understanding of the causes of these trends even more difficult. **Figures B1** and **B2** illustrate reading trends in the United States and the six comparison countries from 2000 to 2003 to 2006 to 2009.

To make the figures easier to understand, we display trends for disadvantaged social class (Groups 1 and 2) students and advantaged social class (Groups 5 and 6) students only. As the previous discussion has made clear, it would not be accurate to assume that the trends for middle social class (Groups 3 and 4) students, not shown, in each case parallel the trends for advantaged and disadvantaged students.

Before reasonable policy conclusions can be based on PISA reading score trends from 2000 to 2009, we should attempt to understand why, in the lowest (Group 1) social class group, reading scores improved substantially for U.S. and German students but declined for U.K. and Canadian students and declined substantially for students in France and Finland. Likewise, we should attempt to understand why reading scores for U.S., German, and Canadian students in the highest (Group 6) social class group were unchanged but improved substantially for comparable social class students in France and Korea and declined for students in Finland and the United Kingdom. We should understand why there was a collapse in reading performance across all social class groups in the United Kingdom, and we should understand why in Korea there was improvement for upper-middle and advantaged social class (Groups 4-6) students only. We are not aware of differing socioeconomic trends or changes in instructional or educational policies that can help to explain these disparate reading results, and so are not persuaded by policymakers who draw conclusions from these test score trends. Simple and seemingly obvious explanations cannot account for these complex results. If curricular or instructional changes are responsible, why should they have affected different social class groups within a country differently? If (in the case of Finland, for example) immigration of less literate families explains the drop in Group 1 scores, how does this explain why Group 6 scores fell as well?

FIGURE B1

Reading scores, by social class group, U.S. compared with similar post-industrial countries, PISA 2000–2009

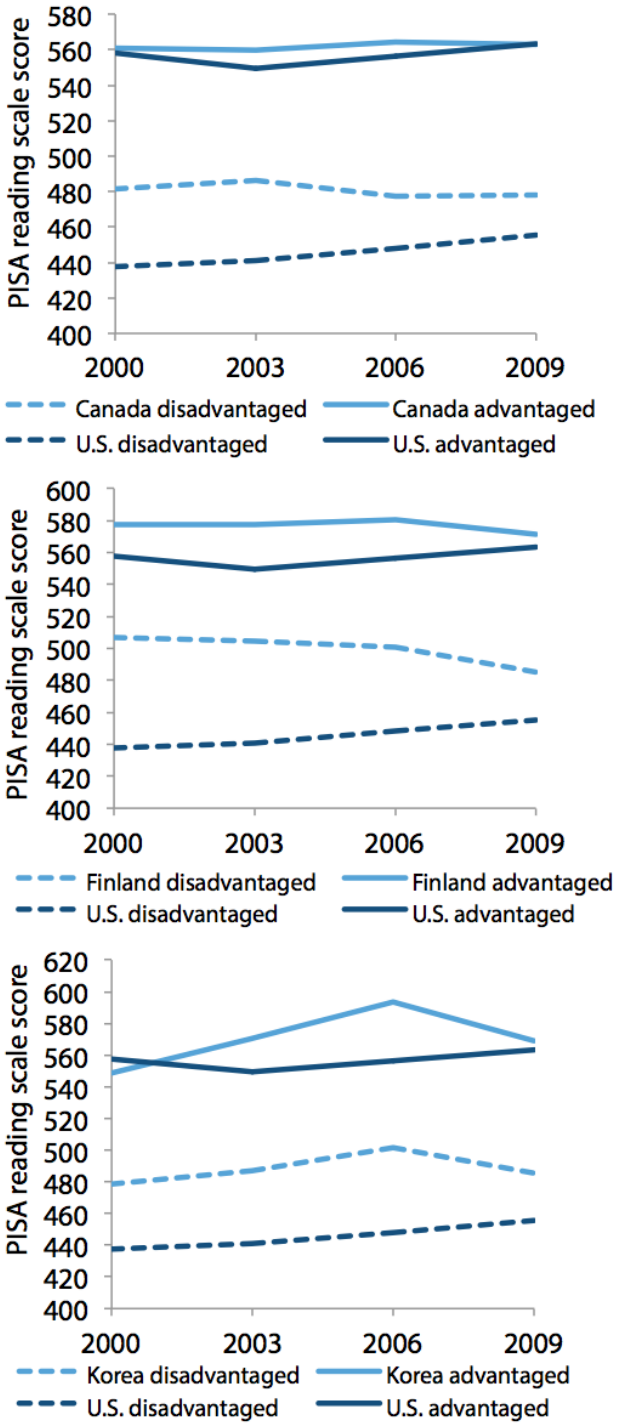


Note: U.S. data for 2006 are unavailable and therefore linearly interpolated.

Source: Authors' analysis of PISA 2000, 2003, 2006, and 2009 databases; authors' calculations of mean test scores by books in the home (BH)

FIGURE B2

Reading scores, by social class group, U.S. compared with top-scoring countries, PISA 2000–2009



Note: U.S. data for 2006 are unavailable and therefore linearly interpolated.

Source: Authors' analysis of PISA 2000, 2003, 2006, and 2009 databases; authors' calculations of mean test scores by books in the home (BH)

TABLE 11A

Mathematics score changes, scale scores by social class group for U.S. and similar post-industrial countries, PISA 2000–2009

	FRANCE			GERMANY			U.K.			U.S.		
	2000	2009	Change	2000	2009	Change	2000	2009	Change	2000	2009	Change
Group 1 (Lowest)	458	413	-45	381	433	52	458	435	-23	416	434	18
Group 2	484	460	-24	418	466	49	483	455	-29	446	464	17
Group 3	517	498	-19	471	509	39	519	487	-32	490	491	1
Group 4	537	529	-8	500	535	36	540	517	-23	510	510	0
Group 5	558	562	4	537	571	34	563	547	-16	543	548	4
Group 6 (Highest)	544	569	24	550	570	20	579	551	-29	554	548	-6
National average math score	517	497	-20	490	513	23	529	492	-37	493	487	-6

Source: Authors' analysis of OECD Program for International Student Assessment (PISA) 2000 database, with authors' interpolations of average test scores, and Tables 1 and 4 for 2009 data

As noted above, socioeconomic, instructional, or educational changes anywhere from 1985 (the birth year of students taking PISA in 2000) through 2009 could help explain these changes in performance of 15-year-olds over the nine years from 2000 to 2009. Complicating matters further, average scores for countries, or for separate social class groups, did not trend in a straight line from 2000 to 2009. In some cases an overall increase was the consequence of a drop during one interim period but a larger gain in another. Attempting to explain changes in performance over interim periods, however, would be even more difficult than attempting to explain them over the full nine years.

Mathematics, 2000–2009

Table 11A displays how math achievement changed from 2000 to 2009 in the United States and the three similar post-industrial countries.

Table 11B displays data on how math gaps between U.S. students and comparable social class students in the three similar post-industrial countries changed from 2000 to 2009. (Positive numbers describe gains for U.S. performance relative to the performance of comparison countries. Negative numbers describe deteriorated U.S. performance relative to that of comparison countries.)

TABLE 11B

Math score gap changes, U.S. vs. similar post-industrial countries, PISA 2000–2009

	GAP CHANGES, U.S. VERSUS:		
	France	Germany	U.K.
Group 1 (Lowest)	+63	-35	+41
Group 2	+41	-31	+46
Group 3	+20	-38	+33
Group 4	+8	-35	+23
Group 5	+1	-29	+21
Group 6 (Highest)	-30	-27	+22

Note: Numbers in this table take the 2009 U.S. average score for a social class group, less the 2009 comparison country's average score for the same social class group, and subtract from this result the 2000 U.S. average score for that social class group, less the 2000 comparison country's average score for the same social class group.

Source: Table 11A

Considering the full 2000 to 2009 period, U.S. math scores improved for disadvantaged social class (Group 1 and 2) students, and were unchanged for the four other social class groups. These changes are very similar to the changes described above for reading from 2000 to 2009

(Table 9A); in the case of both reading and math, the gains for Group 1 improved substantially and by a similar amount.

Considering the three similar post-industrial countries in this full 2000 to 2009 period:

- In France, math scores declined substantially for disadvantaged and for lower-middle-class students (Groups 1-3), declined for upper-middle-class students (Group 4), were stagnant for higher social class students (Group 5), and improved substantially for the highest social class students (Group 6). Overall, these changes in French math performance from 2000 to 2009 were similar in direction to the changes in reading performance over the same period, although in the bottom social classes the deterioration in math performance was much more severe than in reading performance. Also similar to reading, although disadvantaged social class (Groups 1-2) students in the United States scored substantially worse than comparable social class French students in math in 2000, by 2009 these U.S. students scored better than their French social class counterparts and substantially better in the lowest social class (Group 1). However, and again as in reading, whereas in 2000 U.S. students from the highest social class (Group 6) scored above French students, by 2009 these students scored below their French social class counterparts.
- In Germany, math scores improved substantially in every social class group, with especially large gains for the lowest social class (Group 1) students. As in France, these trends in math were very similar to the changes described in reading (Table 9A), the only exception being that German reading scores for the highest social class (Group 6) students were unchanged from 2000 and 2009, but in math these students made large gains, nearly as large as those made by German students in other social class groups. In 2000, German disadvantaged and middle-class students (Groups 1-4) scored worse in math than comparable social class U.S. students, while German advantaged social class (Groups 5-6) students scored about the same in math as their U.S. counterparts. But in 2009, U.S. and German disadvantaged social class (Groups 1-2) students scored about the same, while German middle and advantaged social class students (Groups 3-6) now scored substantially better than comparable social class U.S. students. Although the direction of these relative changes was similar in math and reading, their magnitude was much greater in math than in reading.
- In the United Kingdom, math scores declined in every social class group, with substantial declines for disadvantaged, middle-class, and the highest social class (Groups 1-4 and 6) students. As in the United States, France, and Germany, these trends in math were very similar to the changes in reading (Table 9A) from 2000 to 2009.

Table 12A displays how math achievement changed from 2000 to 2009 in the United States and the three top-scoring countries.

Table 12B displays data on how math gaps between U.S. students and comparable social class students in the three similar post-industrial countries changed from 2000 to 2009. Positive numbers describe gains for U.S. performance relative to the performance of comparison countries. Negative numbers describe deteriorated U.S. performance relative to that of comparison countries.)

Considering the three top-scoring countries in this full 2000 to 2009 period:

- In Canada, math scores declined for disadvantaged (Groups 1 and 2) students, remained about the same for middle-class students (Groups 3 and 4), and improved for advantaged students (Groups 5 and 6). Despite declines at the bottom of the social class distribution, Canadian students still scored above U.S. students in every social class group in 2009, although the Canada-U.S. gap narrowed for disadvantaged

TABLE 12A

Mathematics score changes, scale scores by social class group for U.S. and top-scoring countries, PISA 2000–2009

	CANADA			FINLAND			KOREA			U.S.		
	2000	2009	Change	2000	2009	Change	2000	2009	Change	2000	2009	Change
Group 1 (Lowest)	487	471	-17	507	490	-17	473	452	-20	416	434	18
Group 2	502	493	-9	518	507	-11	501	504	3	446	464	17
Group 3	523	521	-1	527	528	1	538	531	-8	490	491	1
Group 4	539	543	4	544	552	8	556	553	-3	510	510	0
Group 5	551	560	9	558	570	12	577	579	2	543	548	4
Group 6 (Highest)	556	567	11	565	580	15	588	602	13	554	548	-6
National average math score	533	527	-6	536	541	4	547	546	-1	493	487	-6

Source: Authors' analysis of OECD Program for International Student Assessment (PISA) 2000 database, with authors' interpolations of average test scores, and Tables 1 and 5 for 2009 data

TABLE 12B

Math score gap changes, U.S. vs. top-scoring countries, PISA 2000–2009

	GAP CHANGES, U.S. VERSUS:		
	Canada	Finland	Korea
Group 1 (Lowest)	+34	+35	+38
Group 2	+26	+28	+15
Group 3	+2	0	+9
Group 4	-3	-8	+3
Group 5	-5	-7	+2
Group 6 (Highest)	-17	-21	-19

Note: Numbers in this table take the 2009 U.S. average score for a social class group, less the 2009 comparison country's average score for the same social class group, and subtract from this result the 2000 U.S. average score for that social class group, less the 2000 comparison country's average score for the same social class group.

Source: Table 12A

social class (Groups 1 and 2) students and widened for the highest social class (Group 6) students during this period.

- In Finland, math trends were very similar to those in Canada: scores declined for disadvantaged social class (Groups 1 and 2) students, were unchanged for lower-middle social class (Group 3) students, and improved for upper-middle and advantaged social class (Groups 4-6) students. The math score decline for Finland's disadvantaged social class (Groups 1-2) students was not as great as its reading score decline for comparable social class students. Finland's math scores improved for its upper-middle and advantaged social class (Groups 4-6) students from 2000 to 2009, an improvement not seen in reading for these students. As in the case of Canada, although the United States continued to score below Finland in each social class group, this negative test score gap narrowed for disadvantaged students but widened for the highest social class (Group 6) students.
- In Korea, as in Canada and Finland, math scores declined for the lowest social class (Group 1) students, and in Korea the decline was substantial. Korean math scores improved for the highest social class (Group 6) students from 2000 to 2009, with groups in between these bottom and top groups remaining about the same.

Thus, there was a narrowing math gap from 2000 to 2009 between disadvantaged U.S. students and comparable students in each of the top-scoring countries and a widening math gap between the highest social class students in the United States and comparable students in each of the

top-scoring countries. Disadvantaged students in the top-scoring countries, however, continued to outperform disadvantaged students in the United States in math in 2009, though by a smaller margin.

Comparing trends in both reading and math for the full 2000 to 2009 period for the United States and the three top-scoring countries, there was a narrowing of the gap for disadvantaged students. There was a widening of the gap for students at the top of the social class scale in both reading and math between the United States and Korea. For Canada and Finland, however, although the gap narrowed for disadvantaged students in reading and math, it widened for the highest social class (Group 6) students in math but not in reading.

Following release of the PISA 2009 scores, U.S. policymakers and critics have devoted considerable attention to education in Finland because PISA 2009 scores in Finland were considerably higher than those in the United States, both in reading and math. However, less attention has been paid to the fact that, as we have shown in these tables, although U.S. 2009 scores are systematically lower than those in Finland, over the last decade U.S. scores for the lowest social class students (Group 1) have improved in both reading and math, while scores for these students in Finland have plummeted. Indeed, scores in Finland fell for disadvantaged social class students (Groups 1 and 2) in both reading and math, and also fell in reading for students from every other social class group except the higher social class (Group 5) students.

When reviewers of this report saw this finding, several speculated that the decline of disadvantaged students' scores in Finland may be attributable to an influx of poorly educated immigrants. But this is unlikely to explain much of the decline in Finland's overall average reading scores, for two reasons. First, as noted, reading scores also declined in Finland for middle and the highest social class (Groups 3, 4, and 6) students. And second, the share of Finnish students who are the most disadvantaged (those in the lowest social class, Group 1) declined from 2000 to 2009 (Table 8A). Of course, it is possible that new immigrants in the disadvantaged social classes performed much worse than other disadvantaged students, and their lower scores more than offset their smaller proportion in the national average. But this is purely speculative, and reinforces the point that before uncritically accepting Finland as an educational model, scholars should look not only at score levels but at score changes, to see if socioeconomic, curricular, or other school changes had an adverse effect.

Policymakers and critics might also learn from 2000–2009 trends in Germany. German average 2009 PISA scores are now about the same as U.S. scores in reading and substantially higher than U.S. scores in math because German scores have improved at a very rapid rate for almost all social class groups in reading (Table 9A) and math (Table 11A) during the last decade.

Yet, scholars have also failed to investigate rigorously the causes of these dramatic improvements in German scores. In a report to U.S. Education Secretary Duncan (OECD 2011), OECD educational experts listed a number of reforms that the German federal government and German states implemented in the wake of Germany's relatively low performance on the 2000 PISA test. These include the beginnings of changes in Germany's highly class-based secondary school structure, longer school days, national standards, and greater school accountability. However, the report notes that "the reforms have been only partially implemented so far and have not yet had time to affect the performance of students who were 15 in 2009" (p. 213). Thus, the reforms would not be able to explain German students' increasing scores from 2000 to 2009.

More than one reviewer of this report suggested that part of the explanation for German improvement might be that the unified country has made unusually large investments in the living standards and educations of residents in the former East Germany, in an attempt to bring them up to living standards and educational quality in the former West Germany. But this superficially plaus-

ible explanation may not be the most important. A careful analysis of German PISA gains in reading seems to show that immigrant groups accounted for much of this score from 2000 to 2009. Ethnic Germans only recorded a small increase (5 points in reading), while immigrant youth had very substantial increases (33 points among first-generation immigrants and 26 points among the second generation). These gains were spread across social class groups but tended to be larger in more disadvantaged youth (Stanat, Rauch, and Segeritz 2010). Whether German rates of improvement among first- and second-generation immigrants, or perhaps among students residing in the former East Germany, can be sustained to the point where Germany's scores by social class group uniformly exceed those of the United States remains to be seen. As of 2009, this was not the case.

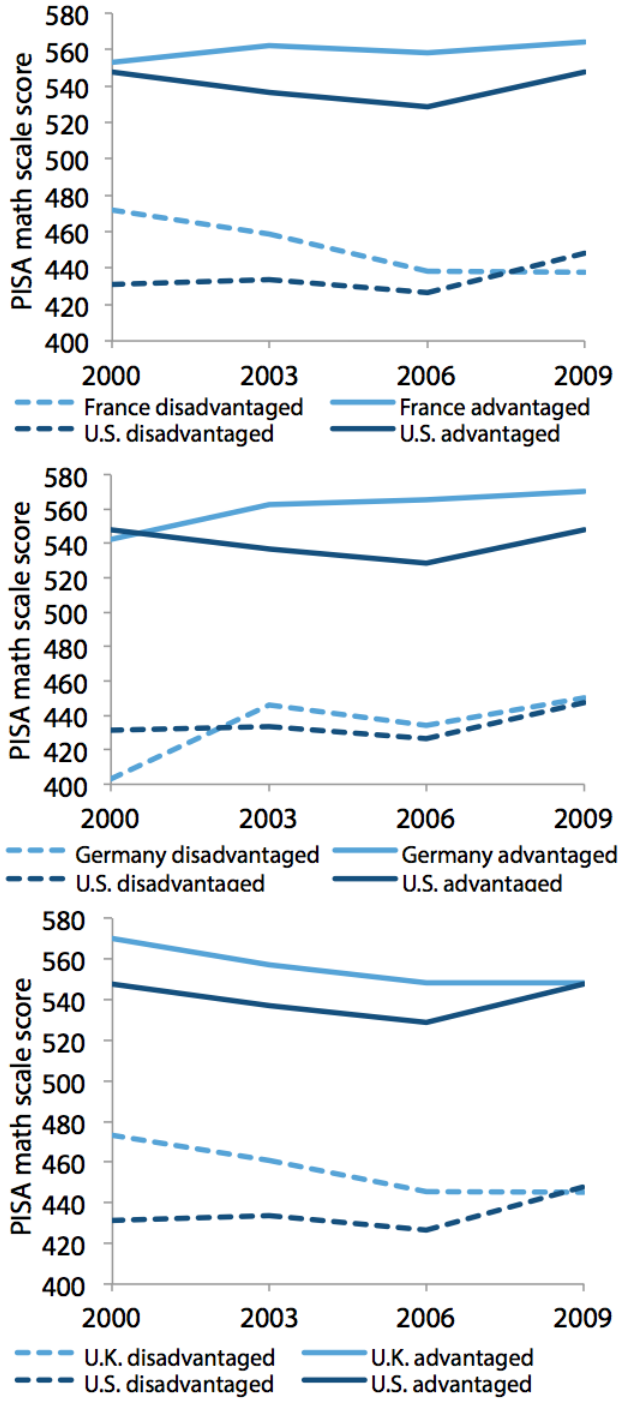
U.S. policymakers should focus their attention in particular on whatever socioeconomic trends, mathematics instructional policies, and educational policies in Germany may have differed from those in the United States during the last decade and in the years leading up to it. Especially noteworthy is that the gap widened between U.S. and German students in every social class group in both reading and math (except for the highest social class students in reading, for whom the gap was mostly unchanged).

Specific curricular instructional changes in both reading and math in Germany could have contributed to these changes. Either socioeconomic or educational policy changes affecting both reading and math performance in the years leading up to 2009 could also have contributed.

Also of note is that the United States and Germany were the only nations in our study whose math and reading performance improved for the lowest social class (Group 1) students from 2000 to 2009. The math and reading performance of the lowest social class students in Canada, Finland, Korea, France, and the United Kingdom all deteriorated in this period.

FIGURE C1

Math scores, by social class group, U.S. compared with similar post-industrial countries, PISA 2000–2009

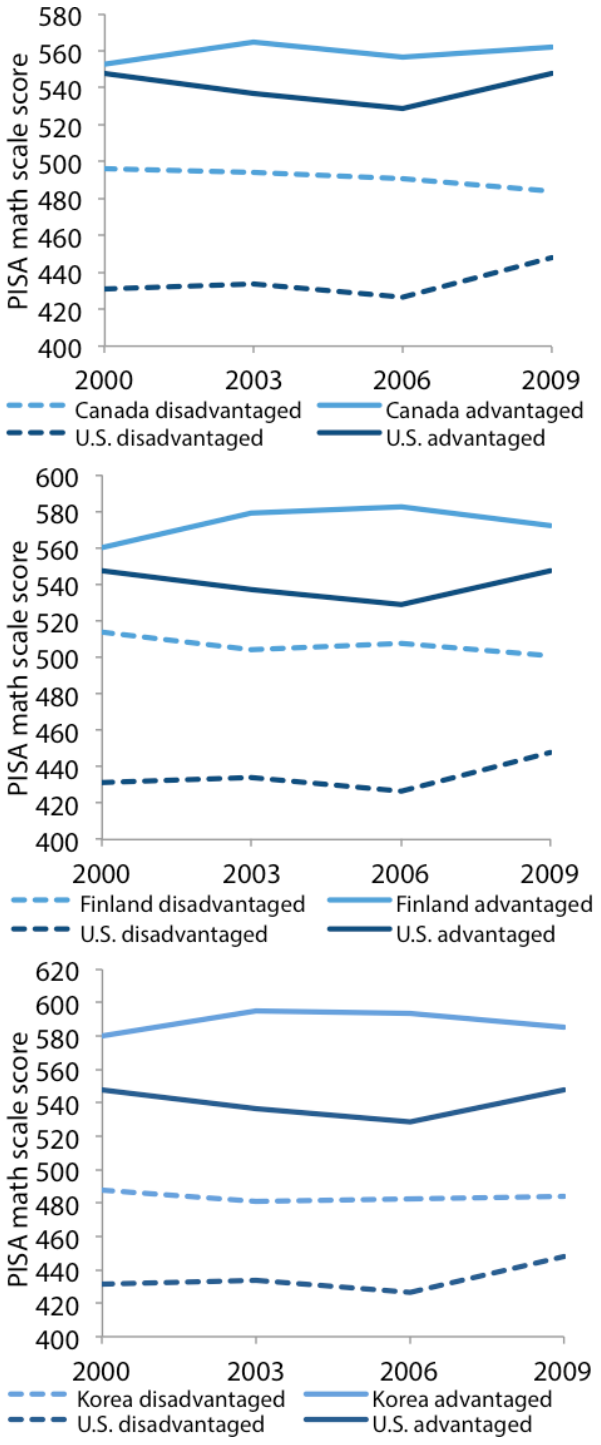


Source: Authors' analysis of PISA 2000, 2003, 2006, and 2009 databases; authors' calculations of mean test scores by books in the home

As was the case in reading, average math scores for countries, or for separate social class groups, did not trend in a straight line from 2000 to 2009. In some cases an over-

FIGURE C2

Math scores, by social class group, U.S. compared with top-scoring countries, PISA 2000–2009



Source: Authors' analysis of PISA 2000, 2003, 2006, and 2009 databases; authors' calculations of mean test scores by books in the home (BH)

all increase was the consequence of a drop during one interim period but a larger gain in another. Attempting to explain changes in performance over interim periods, however, would be even more difficult than attempting to explain them over the full nine years.

Figures C1 and **C2** illustrate math trends in the United States and the six comparison countries from 2000 to 2003 to 2006 and to 2009.

As with the reading figures (Figures B1 and B2), Figures C1 and C2 display trends for disadvantaged social class (Groups 1 and 2) students and advantaged social class (Groups 5 and 6) students only. As the previous discussion has made clear, it would not be accurate to assume that the trends for middle social class (Groups 3 and 4) students, not shown, in each case parallel the trends for advantaged and disadvantaged students.

We can attempt to determine the extent to which changes in countries' social class composition from 2000 to 2009 can explain changes in those countries' average PISA performance and what the underlying trend is. **Table 13** reweights PISA 2009 reading and math scores by 2000 social class composition. **Figures D1** (for reading) and **D2** (for math) illustrate this reweighting.

Table 13 shows what the overall national average scores of the United States and six comparison countries would have been in 2009 if the social class composition of each of these countries had not changed subsequent to 2000, and if the average performance of students in each social class in 2009 remained as we have reported it in Tables 9A, 10A, 11A, and 12A.[13] In Table 13, rows f and p show the change in national scores attributable to change in social class composition between 2000 and 2009. Rows g and q show the change attributable to educational improvement (deterioration) or to other factors.

In Figures D1 and D2, the left bar for each country shows the reading and math scores for 2000, recalculated with 2000 social class weights. The right bar shows the actual national average PISA score for 2009. The middle bar shows a recalculation of scores for 2009 by substituting

TABLE 13

The effect of social class compositional changes on test score changes, U.S. and comparison countries, PISA 2000–2009

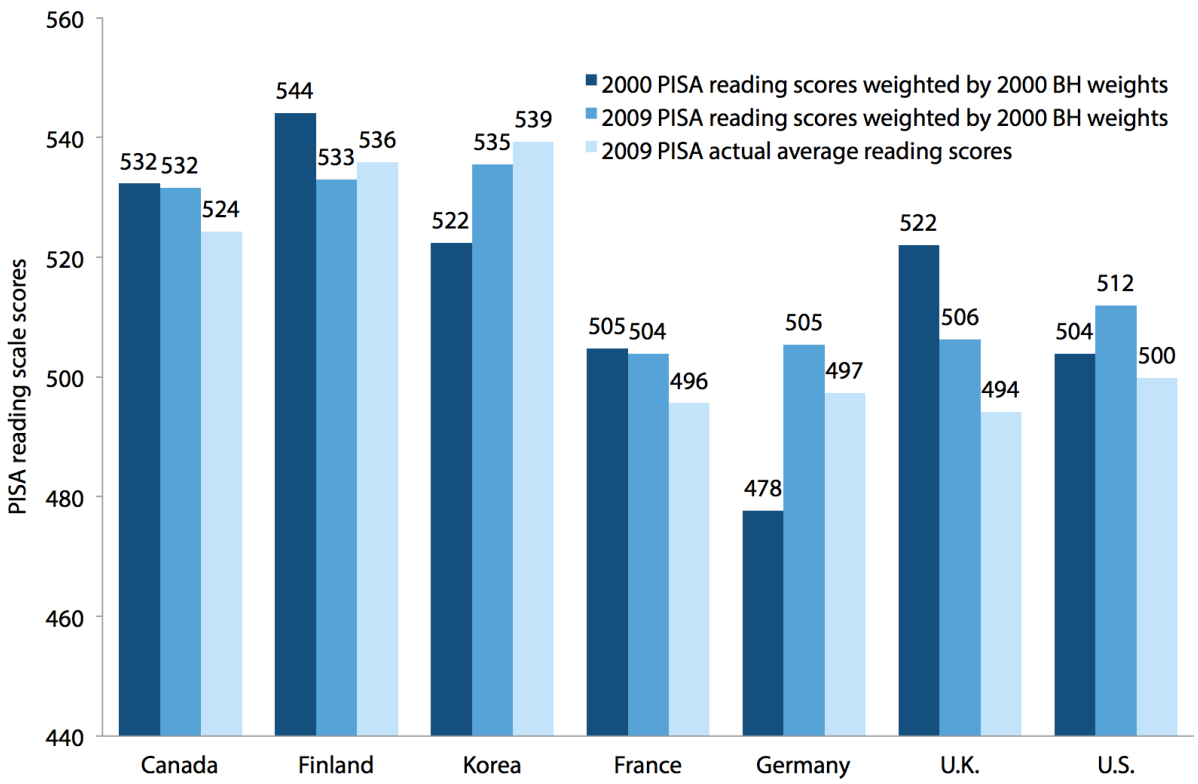
		Canada	Finland	Korea	France	Germany	U.K.	U.S.
Reading								
a	*2000 PISA average reading score*	534	546	525	505	484	523	504
b	*2000 social class weighted average reading score (average of 2000 social class group reading scores, weighted by 2000 social class relative size)*	532	544	522	505	478	522	504
c	*Difference (a-b), see note*	2	2	2	0	6	1	1
d	*2009 PISA average reading score*	524	536	539	496	497	494	500
e	*2009 average reading score, weighted by 2000 social class distribution (average of 2009 social class group reading scores, weighted by 2000 social class relative size)*	532	533	535	504	505	506	512
f	*Change attributable to social class composition change:*							
	Difference, d-e, 2009 average reading score vs. 2009 average reading score with 2000 social class weights	-7	3	4	-8	-8	-12	-12
g	*Change attributable to educational improvement (or deterioration) or to other factors:*							
	Change in average reading scores, 2000-2009, standardized for 2000 social class composition (e-b)	-1	-11	13	-1	28	-16	8
h	*Reported change (sum of social class and educational improvement/other factors):*							
	Change in reported average reading scores, (adjusted reported) 2000- (reported) 2009 (d-b)	-8	-8	17	-9	20	-28	-4
Math								
k	*2000 PISA average math score*	533	536	547	517	490	529	493
l	*2000 social class weighted average math score (average of 2000 social class group math scores, weighted by 2000 social class relative size)*	532	535	544	518	483	529	493
m	*Difference (k-l), see note*	1	1	3	-1	6	1	0
n	*2009 PISA average math score*	527	541	546	497	513	492	487
o	*2009 average math score, weighted by 2000 social class distribution (average of 2009 social class group math scores, weighted by 2000 social class relative size)*	533	538	541	504	521	503	498
p	*Change attributable to social class composition change:*							
	Difference, n-o, 2009 average math score vs. 2009 average math score with 2000 social class weights	-7	3	6	-7	-8	-10	-11
q	*Change attributable to educational improvement (or deterioration) or to other factors:*							
	Change in average math scores, 2000-2009, standardized for 2000 social class composition (o-l)	2	2	-3	-14	37	-26	5
r	*Reported change (sum of social class and educational improvement/other factors):*							
	Change in reported average math scores, (adjusted reported) 2000- (reported) 2009 (n-l)	-5	5	2	-21	29	-36	-6

Note: The differences in rows (c) and (i) could be the result of some test takers not answering the books-in-the-home (BH) question, or of imprecision in our estimation of relative BH weights for 2000.

Source: Data for rows a and d come from tables 9A and 10A; for rows k and n from tables 11A and 12A; and for rows b, e, l, and o from authors' analysis of OECD Program for International Student Assessment (PISA) 2000 and 2009 databases, weighting test scores in each social class category by 2000 percentage in each social class category.

FIGURE D1

The effect of social class composition changes on reading test score changes, U.S. and six comparison countries, PISA 2000–2009



Note: BH is books in the home.

Source: Table 13, Reading

2000 social class weights for the actual 2009 social class weights. For each country, the change from the left bar to the middle bar shows how the social-class-weighted average scores would have changed from 2000 to 2009 if the social class distribution were unchanged during that period. This represents the change in national PISA scores attributable to educational improvement (or deterioration) or to other factors.

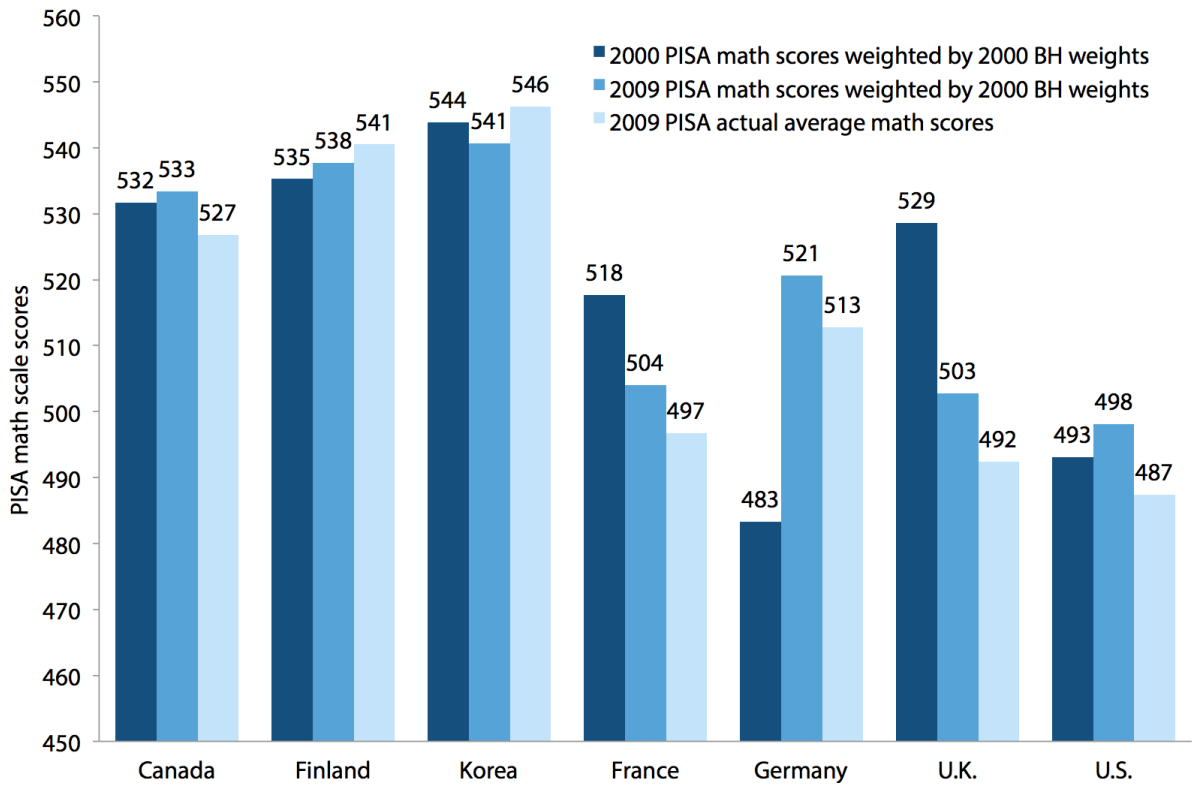
For example, we know from Table 8A that in Canada, the social class composition of Canadian test takers may have deteriorated somewhat from 2000 to 2009. The share of the highest social class (Group 6) test takers declined by 4 percentage points while the share of the lowest social class (Group 1) test takers increased by 2 percentage points. Table 13 shows that the changing social class composition of Canadian test takers from 2000 to 2009 is associated with a decline of 7 points in Canadian reading and math test scores. If Canadian test takers in 2009 had the same social composition as they had in 2000, the average Canadian reading score would have been 532, not the actual average of 524, a difference we consider "about the same."

However, for the post-industrial countries, including the United States, a changing social class composition was associated with a real difference in national average scores.

In France, for example, reported average reading scores declined by 9 scale points from the "calculated" (using 2000 social class weights) 2000 score, but Table 13 shows that we could reasonably have expected the deteriorating social class composition alone of French test takers to be associated with an 8 scale point decline. This suggests that French policymakers should be cautious about assuming

FIGURE D2

The effect of social class composition changes on math test score changes, U.S. and comparison countries, PISA 2000–2009



Note: BH is books in the home.

Source: Table 13, Mathematics

that the decline in test scores has its origin in a failure of educational practice. It is possible that the decline in test scores occurred despite no deterioration in educational practice, which, in the absence of deteriorating social class composition, would have led to no change (-1 scale point) rather than to a decline in scores.

Table 13 suggests a different possibility for the United States. Actual reported reading scores were about the same in 2009 as in "calculated" 2000 (a 4 scale point decline from 504 to 500, or, in other words, scores about the same), but the deteriorating social class composition of U.S. test takers (see Table 8A) suggests a decline of 12 scale points in test scores. That this did not occur suggests the possibility that improved educational practice overcame the harmful effects on overall achievement of the social class compositional changes.

In general, Table 13 and Figures D1 and D2 show that the impact of social class compositional changes for each country on math scores was nearly the same as its impact on reading scores (compare rows f and p). However, the impact of educational or other factors was in some countries very different in reading and math (compare rows g and q). For example, educational and other factors appear to have had a negative effect on reading in Finland but none in math, a positive effect on reading in Korea but none in math, a negative effect on math in France but none in reading, and a positive effect on reading in the United States but none in math. Such differences may be due to specific education policies in each country or to cultural shifts, changes in test curricular coverage, or other factors.

Part IV: Defining social class for comparative purposes

This report emphasizes that policymakers can be led astray if they examine only average national scores without disaggregating those scores by social class and looking at trends, not only levels. However, there is no generally accepted method for classifying students by their social class background. In the United States, analysts typically divide students into only two groups, those who receive full or partial subsidies for school lunches, and those who do not. Since the lunch program and its eligibility requirements are idiosyncratic to the United States, it is neither possible nor desirable to use this criterion to make comparisons between countries. For purposes of this report, we use a categorical index of books in children's homes, with our assumption being that children with fewer books in the home are more socioeconomically disadvantaged with regard to their home preparation for school achievement. In this Part IV, we defend our use of books in the home as the most appropriate of the available measures for analyzing test score differences by students' social class, and we report on its advantages and robustness when compared to other measures of social class, such as mother's education, parents' highest level of education, and the OCED index of social class (the index of Economic, Social, and Cultural Status, or ESCS).

In Parts II and III of this report, we disaggregated the PISA scores of students in the United States and comparison countries by social class group, dividing the test takers into six such groups, from the lowest (Group 1) to the highest (Group 6). We were able to do so because data from PISA are available not only for each 15-year-old student's performance on reading and mathematics tests but also for the student's several socioeconomic characteristics: father's and mother's years of completed schooling; father's and mother's occupational status; and whether the student has a desk at which to study at home, a room of his or her own, a quiet place to study, educational software, a link to the Internet, a personal calculator, classic literature, books of poetry, works of art (e.g., paintings), books to help with school work, a dictionary, a dishwasher, a DVD player or VCR, three other country-specific items, and the number of cellular phones, televisions, computers, cars, and books in the home (Schulz 2005).

It is difficult to know how reliable an indicator of social class many of these characteristics are when considering a cross-national database. Parents' education—particularly, mother's education—is a popular measure of social class within a country because of the likely influence that more-educated parents have on their children's academic attainment and achievement. Yet, using parents' education for cross-country comparisons is more problematic. For example, a country with more universal high school attendance may have higher attainment for relatively lower social classes than countries with less widespread attendance. It is not obvious whether we should consider parents in such countries to have similar socioeconomic status if they have similar educational attainment. We should probably consider the number of a family's cars to have different socioeconomic significance in countries of different geographical sizes and with different transportation infrastructures. Personal computers may have made dictionaries, calculators, and VCRs more obsolete in some countries than in others; we can only speculate about how having a dictionary in one country compares as a social class characteristic to having a personal computer in another country. Perhaps having a personal computer is a more reliable sign of higher social class status in a country where computers are relatively rare, and perhaps having a physical dictionary becomes a less reliable sign of such status in a country where computers are increasing in importance.

In all cases, PISA determines socioeconomic characteristics from a questionnaire completed by students who take the test. Student answers are not always reliable. Parents' education and occupation, for example, are subject to considerable reporting error by 15-year-olds. Nonetheless, using these data, PISA compiles an overall socioeconomic index, the ESCS.

In this report, we do not use the OECD's ESCS index to disaggregate each country's test takers by social class. Instead, we conclude that one element in that index, the number of books in a student's home (hereinafter usually referred to as "BH"), is a more useful and reliable (though still very approximate) indicator by which to make cross-country social class comparisons. This indicator of household literacy is plausibly relevant to student academic performance, and it has been used frequently for this purpose by social scientists (see, for example, Raudenbush et al. 1996).

The ESCS index arbitrarily gives equal weight to parental educational attainment, parental occupational status, and a sub-index of the collection of possessions. Once OECD statisticians calculated the index for each student and weighted the ESCS index by the student weights within each country,[14] they set the mean of the distribution in each country at zero, with a standard deviation of one, and estimated each student's ESCS as the student's standard deviation from the mean of that country's ESCS. The statisticians used the index of student "possessions in the home" to calculate each country's average position relative to the OECD mean and adjusted each student's ESCS index in that country by that constant term. Finally, they combined all the OECD country distributions of ESCS with their adjusted means into a single OECD distribution. To preserve the integrity of country distributions, the statisticians "compressed" the data into an artificial "sample" of one thousand students from each country to construct the distribution of ESCS for the OECD, with a mean of zero and standard deviation of one. The ESCS ranks the index number of each test taker, in all countries, on that single continuous standardized scale. Since each country is given equal weight in constructing the distribution, relative to the number of 15-year-olds in each country, the ESCS of students in smaller countries is weighted more heavily than that of students in larger countries.

Although the methods used to construct the PISA ESCS scale may make sense to statisticians, one illustration of how difficult it is to interpret the scale is that the United States ranks relatively high on the scale, but this is largely attributable to U.S. parents having relatively high educational attainment (years of school completed) in 2009 and a high index of articles in the home. The parents of PISA test takers—15-year-olds—in 2009 would mostly have been between the ages of 40 and 50 at that time; they would have been of college-going age in the 1980s. During this period, the rate of U.S. college attendance was considerably higher than rates in comparable countries, due partly to a U.S. higher education open admissions policy that did not then have a parallel in comparable countries. We do not believe that a U.S. college dropout at that time necessarily had meaningfully higher social class status than a high school graduate in other OECD countries. But because PISA includes parental years-of-school-completed as a cross-national indicator of social class status, the child of such a college dropout would have a higher ESCS rank, other things equal.

The number of books in a home may indicate greater parental literacy and therefore greater student academic advantage, while many physical articles in the home that are measured in the PISA questionnaire for purposes of constructing the ESCS index may not be good predictors of students' academic advantage. Yet physical articles play a major role in setting each county's average position in the OECD's ESCS distribution. In more industrialized countries, for example, television sets and VCRs may be widespread across all social classes, while it is only in less-industrialized countries that the possession of such physical articles indicates higher social class status. The use of physical articles in the home as an important component of ESCS places students in countries such as Korea much lower on the ESCS scale than students in the United States, and makes it appear that when the ESCS index is used to measure social class, average performance in each Korean social class compares more favorably than is in fact the case with performance in the same social class in comparison countries. Students in the Korean sample have a much higher average level of books in the home

than students in the United States. Arguably, books in the home may contribute more to school success than television sets or VCRs, so the small weight of BH in the ESCS index, relative to other physical articles, may make the ESCS index inappropriate for making predictions about academic performance. We discuss this and other issues regarding our choice of BH rather than ESCS in Appendix B.

Because of such questions regarding the OECD's ESCS, we chose not to use it for purposes of social class comparisons in this report and instead use books in the home. Although ESCS is unique to PISA, BH is also available for TIMSS, so the use of BH makes it possible for us to compare TIMSS and PISA performance by social class group (see Part V).[15]

BH is also more suitable for our analysis because, unlike ESCS, it is not a continuous scale but consists of six discrete groupings created by OECD statisticians, who divided students into the six social class groupings we use in this report.[16] Students in the lowest, least-advantaged group (we refer to it as social class Group 1) have 10 or fewer books in their homes. Students in Group 2 have 11 to 25 books; in Group 3 they have 26 to 100 books; in Group 4 they have 101 to 200 books; in Group 5 they have 201 to 500 books; and students in the highest social class group, Group 6, have more than 500 books in their homes.[17] ESCS and BH are highly correlated.[18]

Because BH is divided into six social class groupings and is not a continuous scale, because of questions we have about the validity of some components of ESCS to predict academic performance, and because BH can be used to compare other tests as well as PISA, we focus our analysis of social class correlates of performance on BH rather than ESCS.

Nonetheless, we acknowledge that there is simply no good way to compare social class across countries. Like parents' education and occupation, BH is also subject to student self-reports that are not fully reliable. To take one example, because students were asked to estimate the shelf space devoted to books rather than to count books themselves, countries in which books tend to be thinner will appear to have students from relatively higher social class backgrounds than will countries where families have the same number of books but where books are thicker. The ratio of paperback to cloth books in Asian countries, for example, is greater than the proportion in the United States. For purposes of this report, therefore, this factor may exaggerate the social class status of Korean students relative to the United States and perhaps other comparison countries.

We conducted two checks on the robustness of BH relative to other measures of social class. In the first, we recalculated the average student PISA reading and mathematics scores we estimated by BH, adjusting the scores for the ESCS index as a reasonable additional measure to capture social class groupings. We concluded that calculations of average test scores by social class that use the ESCS index, with its heavy reliance on physical objects in the home, would not yield superior results to calculations that use BH. In the second, we correlated student test scores on the PISA and TIMSS tests in each country with the BH measure, adding two additional measures of social class—mother's education and parents' highest level of education. These correlations suggest that using either of these measures instead of BH would not change the analysis by social class group. We report on these checks in detail in Appendix B.

Part V. Comparing PISA and TIMSS results in mathematics

The dangers of using national average scores to compare nations' student achievement, without disaggregating those scores by social class or attempting to understand whether longitudinal trends are plausible, are illustrated starkly when we compare adolescent mathematics trends in PISA with those in TIMSS by social class and over time. For example, in the period from 1999–2000 to 2007, PISA showed that,

TABLE 14A

National average mathematics scores, Finland and U.S., TIMSS 1999–2011 and PISA 2000–2009

	1999 TIMSS	2000 PISA	2009 PISA	2011 TIMSS	Change (scale points)	Annual rate of change, earliest to latest score
Finland TIMSS	520			514	-6	-0.1%
Finland PISA		536	541		4	0.1%
U.S. TIMSS	502			509	8	0.1%
U.S. PISA		493	487		-6	-0.1%
Finland-U.S. gap	-19	-43	-53	-5		

Source: Trends in International Mathematics and Science Study (TIMSS) as reported in Mullis et al. (2012); OECD Program for International Student Assessment (PISA) (2010a)

for U.S. students, each social class group's performance was either stagnant or declined.[19] *But for approximately the same period, TIMSS showed that each social class group in the United States improved its performance, and for several groups the improvement was substantial. It would take considerably more investigation than we were able to do for this report to form a judgment about whether the trends reported by TIMSS or PISA, or neither, are accurate. In this section, we compare PISA and TIMSS results in more detail.*

Comparisons between student performance on PISA and TIMSS cannot be made for all countries. A challenge to interpretation of international test performance arises because not all countries participate in both PISA and TIMSS. Indeed, there is fierce competition between the two tests for clients. The United States participated in the PISA math assessment in 2000, 2003, 2006, and 2009, and in 8th-grade TIMSS in 1999, 2003, 2007, and 2011.[20] Of our six comparison countries, only Korea participated in the 8th-grade TIMSS in these years. Finland participated in 1999 and then again in 2011.

The TIMSS scale scores we report in the following cannot be compared directly to PISA scale scores because the scales differ. We can compare performance on the two tests only by looking at trends over time. The rule we use for evaluating changes in TIMSS scores is as follows: We consider 8th-grade math scores that differ by less than 7 scale points to be "about the same." Scores that differ by at least 7 but less than 17 points are "better (or worse)" or "higher (or lower)." Scores that differ by 17 points or more are "substantially better (or worse)" or "substantially higher (or lower)." Seventeen scale points in most cases is equivalent to about 0.2 standard deviations.

Canada participated in TIMSS in 1999, but not subsequently. In that year, Canada's 8th-grade TIMSS scores were substantially higher than those of the United States, but not as high as those in Korea.[21] This is similar to the relative standing of these countries in PISA 2000.

Finland participated in TIMSS in 1999 and then not again until 2011. With great policy attention paid to Finland's superior performance to the United States on PISA, it is important to try to understand whether these PISA results are confirmed by TIMSS. **Table 14A** compares the Finland and U.S. experience in TIMSS from 1999 to 2011. Although the periods are too dissimilar for a direct comparison, Table 14A also displays national average mathematics scores for Finland and the United States on PISA from 2000 to 2009.

Table 14A shows that, whereas in 1999 Finland's national average mathematics score was substantially better than the U.S. national average mathematics score, in 2011 the two countries scored about the same on TIMSS mathematics. Yet on PISA, the very substantial superiority of Finland over the United States in 2000 widened by 2009,

TABLE 14B

Mathematics score comparisons, national average scores, Korea and U.S., TIMSS (8th-graders) and PISA (15-year-olds), 1999/2000–2009

	TIMSS 1999	TIMSS "2009"	Change (scale points)	PISA 2000	PISA 2009	Change (scale points)	TIMSS–PISA rough agreement?
National average, Korea	587	605	18	547	546	-1	No
National average, U.S.	502	509	7	493	487	-6	No

Note: TIMSS "2009" scores are calculated as average of TIMSS 2007 and 2011 score for each country.

Source: Trends in International Mathematics and Science Study (TIMSS) as reported in Mullis et al. (2012); OECD Program for International Student Assessment (PISA) (2010a)

opening a gap that was over half a standard deviation. Does Finland now outperform the United States in adolescent mathematics? The answer seems to depend on which test is cited.

As noted above, the TIMSS 2011 international database, including average scale scores disaggregated by social class, has not yet been released. But we know from our examination of earlier PISA and TIMSS tests that Finland's social class composition is considerably more advantageous than that of the United States. Once the database is released, it will be possible to adjust national average scores in Finland and in the United States for social class composition. Such an adjustment may well show that the United States outperforms Finland overall on TIMSS, once social class composition has been controlled.

A closer comparison can be performed for the United States and Korea, because, unlike Finland, both countries also participated in TIMSS 2007. To see whether TIMSS and PISA trends are similar, **Table 14B** compares national average results for Korea and the United States from 1999 (for TIMSS) and 2000 (for PISA) to 2009 on both PISA and TIMSS. To estimate 2009 performance on TIMSS, the table averages national average results from TIMSS 2007 and TIMSS 2011. We call this constructed result TIMSS "2009." Korea and the United States are the only two countries studied in this report that participated in TIMSS in 1999, 2007, and 2011. None of the three similar post-industrial countries did so.

We can see from the table that trends in the two assessments have not been consistent. Over roughly the same period, Korea's average national score improved substantially on TIMSS, but was unchanged on PISA. The United States improved on TIMSS but was about the same on PISA.[22]

Table 14C compares TIMSS trends from 1999 to 2009 for England with PISA trends from 2000 to 2009 for the United Kingdom. (The United Kingdom as a whole did not participate in TIMSS, and England separately did not participate in PISA.)

The table shows that in England, national average TIMSS scores improved from 1999 to TIMSS "2009." But in approximately the same period, PISA scores in the United Kingdom fell substantially overall, indeed collapsed.

As noted, we cannot examine possible inconsistencies in performance trends of social class groups in this period because the TIMSS 2011 database has not yet been released. Instead, our examination of TIMSS-PISA correspondence by social class can go only to 2007. **Table 15A** compares Korea and the United States from 1999 (for TIMSS) and 2000 (for PISA) to 2007 on both PISA and TIMSS. To estimate 2007 performance on PISA, the

TABLE 14C

Mathematics trends, England and U.K., TIMSS 1999–2009 and PISA 2000–2009

	ENGLAND			U.K.			
	TIMSS 1999	TIMSS "2009"	Change (scale points), 1999–2009	PISA 2000	PISA 2009	Change (scale points), 2000–2009	TIMSS – PISA rough agreement?
National average	496	510	14	529	492	-37	No

Note: TIMSS "2009" scores are calculated as average of TIMSS 2007 and 2011 score for each country.

Source: Trends in International Mathematics and Science Study (TIMSS) as reported in Mullis, et al. (2012); OECD Program for International Student Assessment (PISA) (2010a)

TABLE 15A

Mathematics score comparisons, by social class group, Korea and U.S., TIMSS (8th-graders) and PISA (15-year-olds), 1999/2000–2007

Social class groups	TIMSS 1999	TIMSS 2007	Change (scale points)	PISA 2000	PISA "2007"	Change (scale points)	TIMSS–PISA rough agreement?
Korea							
Group 1 (Lowest)	527	528	1	473	450	-23	No
Group 2	550	548	-2	501	502	1	Yes
Group 3	581	584	2	538	532	-6	Yes
Group 4	605	613	7	556	554	-2	No
Group 5/6 (Higher/ highest)	625	643	18	580	591	11	No
National average	587	597	10	547	547	0	No
United States							
Group 1 (Lowest)	439	461	23	416	423	7	No
Group 2	461	482	22	446	445	-1	No
Group 3	495	515	20	490	479	-11	No
Group 4	523	538	15	510	504	-6	No
Group 5/6 (Higher/ highest)	537	546	8	548	535	-13	No
National average	502	508	7	493	479	-14	No

Note: PISA "2007" is a weighted average of PISA 2006 and PISA 2009 data, where PISA 2006 has twice the weight of PISA 2009.

Source: Trends in International Mathematics and Science Study (TIMSS) as reported in Mullis et al. (2012); OECD Program for International Student Assessment (PISA) (2010a)

table averages national average results from PISA 2006 and PISA 2009, with PISA 2006 weighted twice as heavily as PISA 2009. We call this constructed result PISA "2007."

Although PISA reports books in the home for six social class groups, TIMSS reports for only five. Based on the 2003–2009 categories, Groups 1-4 have identical definitions in the two tests, but TIMSS collapses PISA's two

advantaged social class groups into a single top group of 200 or more books in the home, so, in this table, PISA scores for social class groups 5 and 6 are averaged using the sample proportions in PISA social class Groups 5 and 6 to create a result comparable to TIMSS Group 5.

As with the previous table, cut points for distinguishing improvement from stagnation on TIMSS and PISA differ slightly because the scales on the tests are not identical. By focusing on trends, however, we can see from this table that patterns for Korea and the United States in TIMSS and PISA are dissimilar.

For example, TIMSS suggests that performance of the lowest social class (Group 1) students in Korea was about the same in math in 1999 and 2007. But for roughly the same period, PISA shows the performance of these students falling substantially. TIMSS suggests that the performance of upper-middle social class (Group 4) students improved from 1999 to 2007. But for roughly the same period, PISA shows that the performance of these students was about the same. TIMSS suggests that the performance of advantaged social class (Group 5/6) students was substantially higher in 2007 than in 1999. PISA confirms that the performance of these students was higher, but finds only a modest, not a substantial, improvement. Overall, TIMSS shows improvement for Korean students from 1999 to 2007, while PISA shows that scores were about the same.

For the United States, the TIMSS-PISA differences are even greater. TIMSS suggests that all social class groups improved their math performance from 1999 to 2007, with disadvantaged and lower-middle social class (Groups 1-3) students improving substantially. PISA shows that over roughly the same period, math performance of disadvantaged and upper-middle social class (Groups 1, 2, and 4) students was about the same, but the performance of lower-middle and advantaged social class (Groups 3, 5/6) declined.

For the United States, a decline in U.S. PISA mathematics scores took place from 2000 to 2006 (see Figures C1 and G), but during a similar period (1999 to 2007), U.S. TIMSS mathematics scores improved, and for disadvantaged and lower-middle social class (Groups 1-3) students, they improved substantially. We are aware of no explanation of why U.S. scores should have diverged so much on the two tests during this period.

The last column of Table 15A emphasizes that, in almost all social class groups, there is disagreement between TIMSS and PISA on mathematics performance trends in Korea and the United States over roughly the same time period. Except for lower and lower-middle social class (Groups 2 and 3) students in Korea, in no other case, for either Korea or the United States, do both TIMSS and PISA show changes in scale scores that are about the same, better (or worse), or substantially better (or worse). Such discrepancies raise questions about whether it is ever appropriate to reach conclusions about test score trends by relying on one test or the other.

We can also make similar though less precise comparisons with the United Kingdom because, although it has not itself participated in TIMSS, England participated separately in 1999, 2007, and 2011.

Table 15B compares TIMSS score changes in England from 1999 to 2007, and compares these with changes in U.K. PISA mathematics scores for approximately the same periods. As above, we estimate PISA scores for 2007 by creating a weighted average of PISA scores for 2006 and 2009, where 2006 scores have twice the weight as 2009 scores.

The table shows that in England, national average TIMSS scores improved substantially from 1999 to 2007. Scores improved for each social class group, and substantially for all social class groups except the lowest social class (Group 1) students.

TABLE 15B

Mathematics trends by social class group, England and U.K., TIMSS 1999–2009 and PISA 2000–2009

	ENGLAND			U.K.			
	TIMSS 1999	TIMSS 2007	Change (scale points) 1999–2007	PISA 2000	PISA "2007"	Change (scale points) 2000–2007	TIMSS–PISA rough agreement?
Group 1 (Lowest)	438	452	15	458	437	-21	No
Group 2	456	485	29	483	453	-30	No
Group 3	488	521	34	519	490	-29	No
Group 4	505	536	31	540	514	-26	No
Group 5/6 (Higher/ highest)	537	568	30	570	548	-22	No
National average	496	513	17	529	494	-35	No

Note: PISA "2007" is a weighted average of PISA 2006 and PISA 2009 data, where PISA 2006 has twice the weight of PISA 2009.

Source: Authors' analysis of Trends in International Mathematics and Science Study (TIMSS) 1999 and 2007 databases (Boston College International Study Center) and OECD Program for International Student Assessment (PISA) 2000, 2006, and 2009 databases

But in approximately the same period, PISA scores in England fell substantially overall and for every social class group.

The grossly inconsistent results between PISA and TIMSS trends, displayed in Tables 14A, 14B, 15A, and 15B cannot give us confidence in the reliability of either TIMSS or PISA with regard to changes in adolescents' mathematics performance during the decade of the 2000s.

For England, part of the explanation for the discrepancy may be that, while PISA aggregates results for the United Kingdom as a whole, TIMSS reports scores separately for England and Scotland (Wales and Northern Ireland do not participate in TIMSS), and while England's TIMSS scores increased from 1999 to 2007, Scotland's scores (not shown in the table) declined. This separate reporting, however, cannot fully explain the inconsistency, because England's scores increased more than Scotland's scores declined, and the population of English students is much greater than the population of Scottish students.

Note that in Table 15B, England's countrywide average TIMSS score increased by a substantial 17 points from 1999 to 2007, although all but the lowest social class groups had much greater score increases. The lower overall average gain is a composition effect, attributable to the fact that the 2007 TIMSS sample for England had a considerably worse social class distribution than the 1999 sample. For example, in 1999 only 19 percent of total TIMSS test takers were disadvantaged (Groups 1 and 2), but in 2007 36 percent of total TIMSS test takers were disadvantaged.

A similar apparent anomaly characterizes U.S. TIMSS scores. From 1999 to 2007, the U.S. countrywide average TIMSS score increased by 7 points, but each social class group had a greater increase, except for advantaged students (Groups 5/6), whose scores increased about the same as the overall average. The increase for each of the other social class groups was more than twice or three times the national average increase. This, too, is a composition effect, attributable to the fact that the 2007 TIMSS sample for the United States had a considerably worse social class distribution than the 1999 sample. For example, in 1999, 22 percent of U.S. test takers were disadvantaged (Groups 1 and 2), but in 2007 37 percent of U.S. students were in the disadvantaged social class groups.

This deterioration in the social class composition of the TIMSS sample from 1999 to 2007 is greater than the deterioration in the PISA social class distribution during a roughly similar period. For the United Kingdom in PISA, 9 percent of the sample was from the lowest social class (Group 1) in 2000, and 13 percent was from this social class group in PISA "2007," a small difference, and smaller than the change in the TIMSS sample, where 7 percent were from the lowest social class (Group 1) in 2000, compared to 15 percent in 2007. For the United States, 13 percent of the PISA sample was from the lowest social class (Group 1) in 2000, compared to 17 percent in "2007," an increase of about one-third that is difficult to reconcile with the doubling in the share of the TIMSS sample (from 8 percent to 17 percent) that was categorized in the lowest social class (Group 1) during a roughly similar period.

The large shifts in the TIMSS sample over an eight-year period may be possible, but are suspect in light of the smaller social class distribution shifts in PISA and NAEP. The large TIMSS shifts could be attributable to a change in reporting behavior on the part of students asked to record books in the home, or to TIMSS sampling flaws—sampling an unrepresentatively large number of disadvantaged students and a correspondingly unrepresentatively small number of advantaged students in both England and the United States in 2007, compared to 1999. As we discuss in greater detail below, one reason for caution about these data is that the increase in the share of U.S. TIMSS test takers with few books in the home does not seem consistent with the stability of the share of NAEP test takers over the same period whose mothers had a high school education or less.

Of the 84 percent of U.S. TIMSS test takers who answered the question on mother's education in 1999, 10 percent reported that their mothers had less than a high school degree, about the same as reported in the NAEP. In 2003, the proportion of the TIMSS sample reporting mothers with less than a high school degree increased to 15 percent, and in 2007 to 16 percent. In contrast, in the NAEP sample, the proportion of test takers reporting that their mothers had less than a high school degree remained stable at about 10 percent in the same period. We were also able to estimate the proportion of those students. In the PISA samples, the proportion of test takers with mothers having less than a high school diploma was 9 percent in 2003, 11 percent in 2006, and 11 percent in 2009. These PISA shares are more consistent with NAEP than with TIMSS shares.[23]

Whatever the explanation, these implausible shifts in social class in the TIMSS over such a short period of time provide further reason to treat international test scores with considerable caution and to avoid making policy pronouncements based on superficial score comparisons.

Table 16 explores what we can learn from the participation of U.K. component countries in TIMSS 2007, and how this compares with PISA results for approximately the same period for the United Kingdom and the United States.[24]

TABLE 16

Social class distribution and average math scores, U.K. and its countries, compared with U.S., PISA 2006–2009 and TIMSS 2007

	U.K.		ENGLAND		SCOTLAND		U.S. (PISA)		U.S. (TIMSS)	
	Social class distribution (percent)	Average score	Social class distribution (percent)	Average score	Social class distribution (percent)	Average score	Social class distribution (percent)	Average score	Social class distribution (percent)	Average score
	PISA "2007"	PISA "2007"	TIMSS 2007	TIMSS 2007	TIMSS 2007	TIMSS 2007	PISA "2007"	PISA "2007"	TIMSS 2007	TIMSS 2007
Group 1 (Lowest)	13	437	15	452	22	439	17	423	17	461
Group 2	15	453	21	485	24	469	16	445	20	482
Group 3	30	490	28	521	25	499	28	479	28	515
Group 4	18	514	18	536	14	527	18	504	17	538
Group 5/6 (Higher/ highest)	24	548	18	568	15	540	20	535	18	546

Note: PISA "2007" is a weighted average of PISA 2006 and PISA 2009 data, where PISA 2006 has twice the weight of PISA 2009.

Source: Authors' analysis of Trends in International Mathematics and Science Study (TIMSS) 2007 database (Boston College International Study Center) and OECD Program for International Student Assessment (PISA) 2006 and 2009 databases

The first thing to notice about Table 16 is that TIMSS 2007 reports a social class distribution for both England and Scotland that is more skewed toward disadvantaged students than PISA "2007" reports for the United Kingdom as a whole. Because the other components of the United Kingdom (Wales and Northern Ireland) are not large enough to explain this difference (and, in any event, are unlikely to have fewer disadvantaged students than England and Scotland), this discrepancy is unexplained. It could be attributable to flawed sampling, either for PISA 2006, PISA 2009, or TIMSS 2007, or to the unreliability of student reports of books in the home in one or more of these surveys. It is another reason to make us cautious about taking the results of these assessments too literally.

However, the social class distribution reported for the United States is more similar in TIMSS 2007 and PISA "2007." The contrast in apparent reliability of student reports between the United States and the United Kingdom is troubling, and leads us to wonder what other reliability issues may make international comparisons based on such data inappropriate.

Table 16 also shows that TIMSS 2007 student performance in each social class group is higher in England than in Scotland. U.S. performance in TIMSS 2007 was apparently better across all social class groups than performance in Scotland. U.S. performance in TIMSS 2007 was better than performance in England in the lowest social class (Group 1), substantially worse in the advantaged social class (Group 5/6), and about the same for social class groups in between.

For the United States and Canada, TIMSS sampled enough students to generate statistically reliable national results, but the samples were not large enough to generate results for individual Canadian provinces or U.S. states. Yet although Canada participated in TIMSS only in 1999, two Canadian provinces, Ontario and Quebec, participated in the 8th-grade TIMSS in 2003 and 2007, and one additional province, British Columbia, participated in 2007. And although the United States has participated in TIMSS in each year of its administration, in some years some U.S. states have asked that their TIMSS sample sizes be increased to generate state-level average results.

Table 17 explores what we can learn from the participation of Canadian provinces in TIMSS 2007, and how this compares with PISA results in math for approximately the same period for Canada and the United States.[25]

TABLE 17

Social class distribution and average scores, Canada and its provinces compared with U.S., TIMSS 2007 and PISA "2007"

	CANADA		ONTARIO		BRITISH COLUMBIA		QUEBEC		U.S. (PISA)		U.S. (TIMSS)	
	Social class distribution (percent)	Average score	Social class distribution (percent)	Average score	Social class distribution (percent)	Average score	Social class distribution (percent)	Average score	Social class distribution (percent)	Average score	Social class distribution (percent)	Average score
	PISA "2007"	PISA "2007"	TIMSS 2007	TIMSS 2007	TIMSS 2007	TIMSS 2007	TIMSS 2007	TIMSS 2007	PISA "2007"	PISA "2007"	TIMSS 2007	TIMSS 2007
Group 1 (Lowest)	8	475	8	474	9	460	18	501	17	423	17	461
Group 2	13	497	16	489	15	485	26	515	16	445	20	482
Group 3	31	522	31	517	31	513	32	533	28	479	28	515
Group 4	21	540	22	528	21	519	13	553	18	504	17	538
Group 5/6 (Higher/ highest)	27	559	23	544	24	531	12	567	20	535	18	546

Note: PISA "2007" is a weighted average of PISA 2006 and PISA 2009 data, where PISA 2006 has twice the weight of PISA 2009.

Source: Authors' analysis of Trends in International Mathematics and Science Study (TIMSS) 2007 database (Boston College International Study Center) and OECD Program for International Student Assessment (PISA) 2006 and 2009 databases

Again, scale scores from TIMSS 2007 and PISA "2007" cannot be compared, but we can see from this table that Ontario and British Columbia have very similar social class distributions in TIMSS 2007 to Canada's social class distribution in PISA "2007." The most noteworthy observation about Canada, however, concerns Quebec, with a much larger proportion of disadvantaged social class (Groups 1 and 2) students and a smaller proportion of upper-middle social class (Group 4) and advantaged social class (Group 5/6) students than Ontario, British Columbia, or Canada nationwide. But in each social class group, Quebec students perform better than students in British Columbia, Ontario, and Canada overall. Despite its larger share of lower-scoring disadvantaged students, Quebec's overall average performance is better than Canada's as a whole.

Quebec's social class distribution is similar to that of the United States in TIMSS 2007, and Quebec's students in each social class group also outperform comparable social class students in the United States, substantially so in all cases except upper-middle social class (Group 4) students.

Disadvantaged social class (Groups 1 and 2) and lower-middle social class (Group 3) students in the United States perform about the same on TIMSS 2007 as comparable students in British Columbia, but U.S. upper-middle social class (Group 4) students perform substantially better, and advantaged social class (Group 5/6) students perform better than comparable social class students in British Columbia.

U.S. disadvantaged social class (Groups 1 and 2) students perform worse than comparable social class students in Ontario, while U.S. lower-middle social class (Group 3) and advantaged social class (Group 5/6) students perform about the same on TIMSS 2007 as comparable social class students in Ontario. Upper-middle social class (Group 4) students in the United States perform better than comparable social class students in Ontario.

Based on PISA scores, we classify Canada as a top-scoring country in comparison to the United States. Without TIMSS 2007 data from other Canadian provinces, it is not possible to say with certainty where in Canada we should look to find the cause of this overall superior performance. However, based on data we have, it is at least a possibility that for mathematics, the key can be found in Quebec.

Within the United States, only Massachusetts and Minnesota participated separately (and were also included in the overall U.S. sample) in the 8th-grade TIMSS in 2007.

Table 18, reproducing some data from Table 17, compares TIMSS results for British Columbia, Ontario, Quebec, Massachusetts, and Minnesota.

TABLE 18

Social class distribution and average scores, Canadian provinces compared with U.S states, TIMSS 2007

	ONTARIO		BRITISH COLUMBIA		QUEBEC		MASSACHUSETTS		MINNESOTA		U.S.	
	Social class distribution (percent)	Average score	Social class distribution (percent)	Average score	Social class distribution (percent)	Average score	Social class distribution (percent)	Average score	Social class distribution (percent)	Average score	Social class distribution (percent)	Average score
(1) Group 1 (Lowest)	8	474	9	460	18	501	12	478	10	483	17	461
(2) Group 2	16	489	15	485	26	515	15	509	16	511	20	482
(3) Group 3	31	517	31	513	32	533	27	551	30	528	28	515
(4) Group 4	22	528	21	519	13	553	19	564	21	551	17	538
(5) Group 5/6 (Higher/ highest)	23	544	24	531	12	567	26	587	23	560	18	546
(6) Province (state) average	517		509		528		547		532		508	
(7) Province (state) average (with U.S. weights)	511		503		533		538		526		508	

Source: Authors' analysis of Trends in International Mathematics and Science Study (TIMSS) 2007 database (Boston College International Study Center)

It shows that students in each social class group in Minnesota outperformed comparable students in British Columbia and Ontario. Minnesota performance, compared to that in British Columbia, was substantially better in all social class groups except for lower social class (Group 2) students. Minnesota performance, compared to that in Ontario, was better in all social class groups, and substantially better for lower social class (Group 2) and upper-middle social class (Group 4) students. Minnesota performance is substantially lower than performance in Quebec for the lowest social class (Group 1) students, about the same as performance in Quebec for the lower and middle social class (Groups 2-4) students, and lower than performance in Quebec for advantaged social class (Group 5/6) students.

Massachusetts students in almost every social class group perform substantially better than comparable social class students in the three Canadian provinces. The exceptions are the lowest social class (Group 1) students in Quebec, who perform substantially better than comparable social class students in Massachusetts; upper-middle social class (Group 4) students in Quebec, who perform better, but not substantially better, than comparable social class students in Massachusetts; and the lowest social class (Group 1) students in Ontario and the lower social class (Group 2) students in Quebec, both of whom perform about the same as comparable social class students in Massachusetts.

Minnesota and Massachusetts are relatively high per-capita-income states, with relatively low percentages of low-income minority students, so it might seem that the higher socioeconomic background of students in these states compared to that of the average U.S. student is the main factor in their higher overall average test scores. But Table 17 shows that, except for the lowest social class (Group 1) students in Quebec, students in Massachusetts and Minnesota perform about the same or better than comparable social class students in the three Canadian provinces.

Row 6 of Table 18 displays the published average TIMSS 2007 scores of each province or state. Row 7 of Table 18 reweights the average scores, assuming that each province or state had a social class distribution that was similar to that of the United States nationwide. It shows that adjusting for social class composition makes almost no difference in the overall average scores of these provinces and states. The greatest difference is in the case of Massachusetts, where about one-quarter of its seeming superiority to that of the United States overall can be attributed to its more advantageous social class composition.

Thus, the superior overall performance of students in Massachusetts and Minnesota could be attributable in part to social class differences not identified by the books-in-the-home measure (for example, disadvantaged students in Massachusetts and Minnesota may be less geographically concentrated than comparable students in the United States generally), or to better curriculum or instruction, or to other factors.

We noted above that scores on PISA and TIMSS are not comparable, because the scales on the two tests are different. However, it is possible to estimate what a TIMSS score would be if converted to the PISA scale.[26] Doing this for Massachusetts and Minnesota TIMSS scores enables us to compare the TIMSS performance of social class groups in these states with the performance of comparable social class groups in our three similar post-industrial countries.

Table 19 displays these results. All scores are on the PISA scale, with PISA "2007" scores for countries estimated by a weighted average of PISA 2006 and PISA 2009 scores, with PISA 2006 having twice the weight of PISA 2009, and PISA "2007" social class distributions for countries estimated by a similarly weighted average of PISA 2006 and PISA 2009 social class distributions.

TABLE 19

Social class distribution, Massachusetts, Minnesota, and U.S. compared with similar post-industrial countries, on PISA mathematics scale

	MASSACHUSETTS		MINNESOTA		FRANCE		GERMANY		U.K.		U.S.	
	TIMSS 2007 social class distribution (percent)	Average score, TIMSS 2007 on PISA scale	TIMSS 2007 social class distribution (percent)	Average score, TIMSS 2007 on PISA scale	PISA "2007" social class distribution (percent)	Average score, PISA "2007"	PISA "2007" social class distribution (percent)	Average score, PISA "2007"	PISA "2007" social class distribution (percent)	Average score, PISA "2007"	PISA "2007" social class distribution (percent)	Average score, PISA "2007"
(1) Group 1 (Lowest)	12	459	10	464	13	419	11	421	13	437	17	423
(2) Group 2	15	486	16	488	17	453	13	455	15	453	16	445
(3) Group 3	27	523	30	503	31	496	31	501	30	490	28	479
(4) Group 4	19	534	21	523	18	524	20	531	18	514	18	504
(5) Group 5/6 (Higher/ highest)	26	554	23	530	21	560	26	567	24	548	20	535
(6) State (country) social class-weighted average	520		507		497		509		496		479	
(7) State (country) average (with U.S. weights)	514		503		494		498		491		479	

Note: PISA "2007" is a weighted average of PISA 2006 and PISA 2009 data, where PISA 2006 has twice the weight of PISA 2009.

Source: Authors' analysis of Trends in International Mathematics and Science Study (TIMSS) 2007 database (Boston College International Study Center) and OECD Program for International Student Assessment (PISA) 2006 and 2009 databases

Table 19 displays state or country overall average mathematics scores in two ways. Row 6 shows the average score that would be reported based on the actual sample distribution of that state or country. Row 7 shows the state or country overall average score with a standardized social class distribution—in this case, as if each state or country had the same social class distribution as the United States. With either calculation, Massachusetts and Minnesota outperform the three similar post-industrial countries, in some comparisons substantially.

Table 19 also shows that Massachusetts students perform substantially better in mathematics than comparable social class students in almost every social class comparison with the three similar post-industrial countries. The only social class students in these countries that perform better than comparable social class students in Massachusetts are advantaged social class (Group 5/6) students in Germany. Upper-middle social class (Group 4) students in Massachusetts perform better than comparable social class students in France, but not substantially better. Upper-middle social class (Group 4) students in Germany and advantaged social class (Group 5/6) students in the United Kingdom perform about the same as comparable social class students in Massachusetts.

Disadvantaged social class (Groups 1 and 2) students in Minnesota perform substantially better than comparable social class students in each of the similar post-industrial countries. However, advantaged social class (Group 5/6) students in Minnesota perform substantially worse than comparable social class students in each of the similar post-industrial countries. Performance of middle social class students in Minnesota is more similar to the performance of these groups in the similar post-industrial countries; Minnesota middle social class (Groups 3 and 4) students perform better than comparable social class students in the United Kingdom; upper-middle social class (Group 4) students perform worse than comparable social class students in Germany; and lower-middle social class (Group 3) students in France and Germany, and upper-middle social class students in France perform about the same as comparable social class students in Minnesota.

Part VI. Comparing NAEP, PISA, and TIMSS trends

We can attempt to evaluate the relative reliability of PISA and TIMSS in the United States because we have a third and fourth sampled assessment, the Main NAEP and the Long-Term Trend NAEP (LTT), with which we can also make comparisons. We believe that, because the National Center for Education Statistics pays so much more attention to and devotes so many more resources to the NAEP exams than to the international assessments, if either PISA or TIMSS trends are consistent with NAEP, the more consistent international assessment may be more reliable. But we are nonetheless limited because NAEP does not have a social class measure that is comparable to the BH categories in PISA or TIMSS. And for no other country do we have a similar validity check for PISA or TIMSS, making NAEP a check for the validity of international comparisons of no value.

With regard to curricular coverage, in some respects NAEP may be more similar to PISA, and in other respects more similar to TIMSS, so we can't necessarily conclude that a NAEP-PISA score trend correspondence that is better or worse than a NAEP-TIMSS score trend correspondence provides a definitive explanation for PISA-NAEP inconsistencies. Nonetheless, NAEP trends over the decade we are considering, showing improvement in math, seem to be more consistent with TIMSS than with PISA. In reading, there are no TIMSS data, but PISA and Main NAEP scores show some consistency; U.S. reading scores dropped on both the Main NAEP and PISA from 1999–2000 to 2003–2004, although the PISA decline was much steeper than the NAEP decline. From 2003–2004 to 2008–2009, average U.S. performance improved on both the NAEP tests and the PISA. In this section, we explore what light it is possible to shed on U.S. PISA and TIMSS trends by using data from NAEP.

Since 1990, the United States has administered the Main NAEP in math and reading to a representative sample

of 8th-graders nationwide. Since 1992, individual states have had the option to request a large enough sample to generate state-level results, and since 2003, state-level sampling has been mandatory for all 50 states. Since the early 1970s, the Long-Term Trend NAEP has been administered in math and reading to a representative sample of 13-year-olds nationwide; there is no state-level administration of LTT NAEP. As noted above, the LTT purports to assess a constant set of mathematical skills, while the Main NAEP purports to assess skills that reflect contemporary curriculum and expectations. What this means in practice is that the LTT stresses only basic computational skills, while the Main NAEP has more emphasis on mathematical reasoning, including some constructed response items. The relative emphases of Main NAEP and LTT are similar to the relative emphases of PISA and TIMSS, respectively, although Main NAEP does not place as much emphasis on reasoning as does PISA. Main NAEP is probably more similar to TIMSS in its item coverage, while the LTT has a much greater emphasis than TIMSS on basic skills. Thus, a hierarchy of mathematical reasoning to basic skills is probably something like PISA, TIMSS, Main NAEP, LTT. In reading, the LTT also purports to assess an unchanging set of more basic skills, while the Main NAEP purports to assess more inferential and interpretive reading skills. But it is also the case that the reading skills on the Main NAEP are not as high as the level of skills on the PISA.

Table 20 displays the average reading and math scores for U.S. students nationwide on both the Main NAEP and LTT for all years for which data are available.[27] For Main NAEP, NAEP samples 8th-graders; for the LTT, NAEP samples 13-year-olds.[28] The right-hand columns display the total score change and the average annual percentage change (gain) in scores from the earliest to the most recent date for which data are available.

Although the standard deviation on each NAEP test and administration varies, in general it is about 32 scale points. Thus, it is apparent from Table 20 that the overall reading achievement of 8th-graders (or 13-year-olds) nationwide is about the same as it was when these tests were first given.

In math, however, the story is different. The improvement on both tests has been substantial, with the average annual rate of improvement on the Main NAEP about twice as great as that on the LTT.

Table 21 shows only the LTT (13-year-olds) and Main NAEP (8th-graders) data from 1992 to 2008, the period in which both tests were given.[29]

While the unchanging performance in reading over this period is similar in each test, there were gains in math in each test, with the gains occurring at a considerably more rapid rate on the Main NAEP than on the LTT.

NAEP does not have a social class indicator comparable to the books-in-the-home measure in PISA and TIMSS. NAEP does, however, report data on student characteristics in several categories that generally indicate social class status. One is the federal government's free and reduced price lunch program. Students are eligible for this program if their family incomes are below 185 percent of the federal poverty line. Although this income eligibility level varies by family size, for a family of four it is about 35 percent of the national median income. Another indicator is parent educational level. NAEP collects data on both mother and father parent education levels. Another indicator is race and ethnicity. There is considerable overlap in the United States between "black" and socioeconomic disadvantage.

Table 22 displays NAEP scores on both the LTT and Main NAEP by eligibility for free and reduced-price lunch (FRPL), mother who did not complete high school (Mother < HS), and black race (Af Am). We do not claim that these indicators describe the same students who fall into Groups 1 and 2 on the BH measure in PISA and TIMSS, but only that students who have these character-

TABLE 20

U.S. student mean scores in reading and math, Long-Term Trend and Main NAEP

	READING		MATH	
	LTT	Main	LTT	Main
1971	255			
1975	256			
1978			264	
1980	258			
1981				
1982			269	
1984	257			
1986			269	
1988	257			
1989				
1990	257		270	263
1992	260	260	273	268
1993				
1994	258	260	274	
1996	258		274	271
1998		263		
1999	259		276	
2000				274
2001				
2002		264		
2003		263		278
2004	258		280	
2005		262		279
2007		263		281
2008	260		281	
2009		264		283
2011		265		284
Total change	4	5	17	21
Average annual change	0.04%	0.10%	0.19%	0.37%

Note: Not shown are years in which neither the Long-Term Trend (LTT) NAEP nor the Main NAEP was administered.

Source: Authors' analysis of National Center for Education Statistics' NAEP Data Explorer

istics are, on average, more disadvantaged than the average U.S. student.

TABLE 21

U.S. student mean scores in reading and math, Long-Term Trend and Main NAEP, 1992–2008

	1992	2008	Average annual change
Reading			
LTT	260	260	0.00%
Main	260	263	0.08
Math			
LTT	273	281	0.19%
Main	268	282	0.31

Source: Authors' analysis of National Center for Education Statistics' NAEP Data Explorer

TABLE 22

Relatively disadvantaged U.S. students' mean scores in reading and math, Long-Term Trend and Main NAEP, 1978–2011

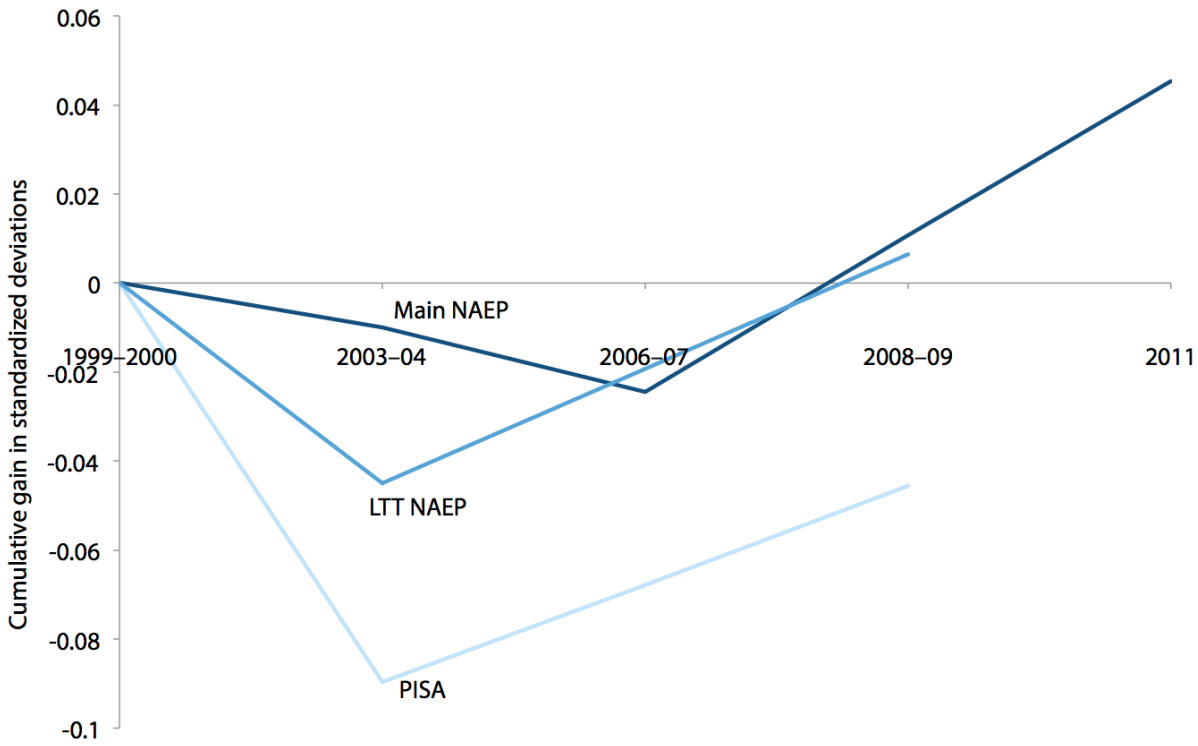
	1978	1980	1982	1984	1986	1988	1990	1992	1994	1996	1998	1999	2000	2002	2003	2004	2005	2007	2008	2009	2011	Total change	Average annual change
Reading																							
LTT																							
FRPL																243			244			1	0.1%
Mother < HS				244		250	246	242	245	243		241				242			243			-2	0.0%
Af Am		233		236		243	241	238	234	234		238				241			247			14	0.2%
Main																							
FRPL											246			249	247		247	247		249	252	6	0.2%
Mother < HS								245	244		248			251	249		247	248		250	250	6	0.1%
Af Am								237	236		243			245	244		243	245		246	249	11	0.2%
Math																							
LTT																							
FRPL																263			266			3	0.3%
Mother < HS					257		257	259		260		261				265			270			13	0.2%
Af Am	230		240		249		249	250	252	252		251				259			262			32	0.4%
Main																							
FRPL										251			254		259		262	265		266	269	18	0.5%
Mother < HS							247	252		255			256		260		262	265		267	268	21	0.4%
Af Am							237	237		241			245		252		255	260		261	262	26	0.5%

Note: FRPL is free and reduced-price lunch; Mother < HS is mother's educational status is less than high school, and Af Am is African American.

Source: Authors' analysis of National Center for Education Statistics' NAEP Data Explorer

FIGURE E

Cumulative gains in Main NAEP, Long-Term Trend NAEP, and PISA reading scores, 1999/2000–2011 (standard deviations)



Note: The data point for PISA in 2006-07 is constructed by linear interpolation.

Source: Authors' analysis of National Center for Education Statistics' NAEP Data Explorer and OECD Program for International Student Assessment (PISA) (2001, 2004, 2007, 2010a)

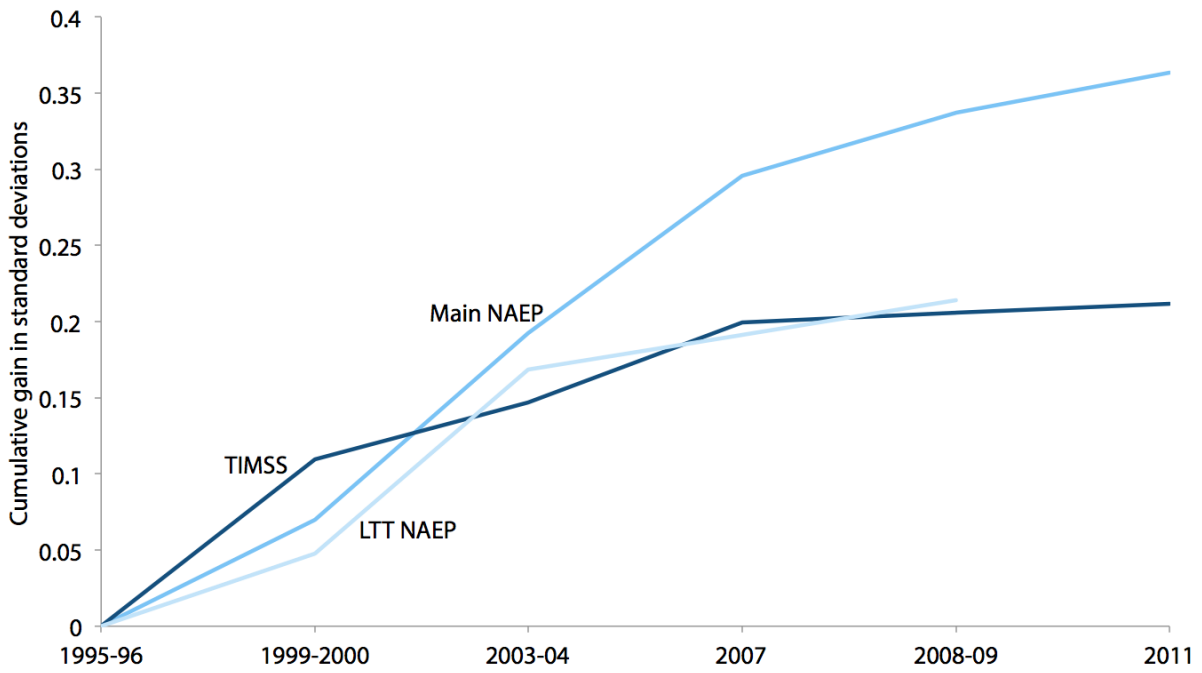
Table 22 shows improvement in reading performance by more disadvantaged students, especially by African American students and especially on the Main NAEP assessment. The relatively greater improvement in reading for more disadvantaged U.S. students than for U.S. students generally (from Table 20) is consistent with what we learned from the PISA reading test.

Figure E compares the U.S. national trends in 8th-grade (13-year-old) reading on the Main NAEP, LTT, and PISA from 1999/2000 to 2009. Figure E also shows the Main NAEP trend through the most recent administration in 2011. Keep in mind in interpreting this and subsequent figures that increases or decreases of about 0.1 standard deviations or less reflect scores that are about the same, increases or decreases of approximately more than 0.1 but less than 0.2 standard deviations are meaningful changes, and increases or decreases of 0.2 standard deviations or more are substantially different.

We are aware of no plausible explanation for the collapse of NAEP and PISA reading scores from 2000 to 2003/2006 and their subsequent recovery in 2009.[30] That the trends were similar in all three tests suggests that the explanation lies not in the design of tests or their administration but in some underlying real characteristic of student performance. If the PISA and the Main NAEP are sampling similar curricula, and if the population samples of the two tests are accurate, Figure E suggests that PISA reading scores in the United States should also increase at the next PISA administration.

FIGURE F

Cumulative gains in Main NAEP, Long-Term Trend NAEP, and TIMSS mathematics scores, 1995/1996–2011 (standard deviations)



Note: The LTT NAEP data point for 2007 is constructed through weighted linear interpolation of the 2003–04 and 2008–09 data points in order to compare to TIMSS 2007.

Source: Authors' analysis of National Center for Education Statistics' NAEP Data Explorer and Trends in International Mathematics and Science study (TIMSS) (Harmon et al. 1997; Mullis et al. 2001, Mullis et al. 2004, Mullis et al. 2008, and Mullis et al. 2012)

Figure F compares the U.S. national trends in 8th-grade math on the Main NAEP, LTT, and TIMSS from 1995 to 2011.

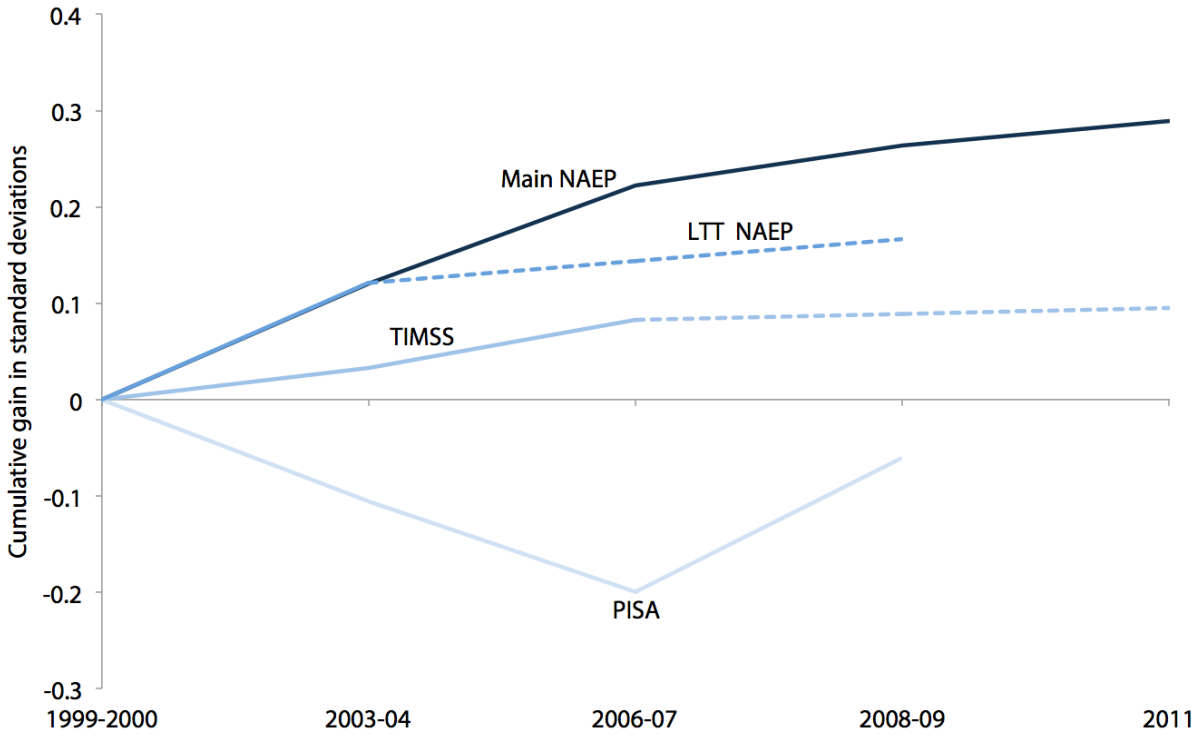
Although the LTT performance from 2004 to 2008 is flat, the direction of overall U.S. national trends in mathematics on the Main NAEP, LTT, and TIMSS from 1995 to 2011 is mostly consistent. The common trend from 1995 to 1999/2000 is especially noteworthy because the social class composition of test takers from 1995 to 1999/2000 was relatively unchanged. It was after 2000 that the share of disadvantaged students in the total test-taking pool began to increase. PISA data are not shown in Figure F because we do not have PISA data prior to 2000.

Figure G begins with 1999/2000, and adds PISA data. It compares the U.S. national trends in 8th-grade (13-year-old) math on the Main NAEP, LTT, PISA, and TIMSS from 1999/2000 to 2011.

As in reading, the collapse of U.S. PISA scores in 2003 does not seem to be replicated in any of the other tests we are considering. U.S. PISA math scores then collapsed further in 2006. Neither the 2000 to 2003 U.S. decline, nor the 2003 to 2006 decline, is replicated in the other test displayed in Figure G. Yet from 2006 to 2009, U.S. PISA math performance increased more rapidly than Main NAEP or TIMSS math performance, both of which remained about the same. We know of no plausible explanation for these apparent trends; the most likely assumption is that the math curriculum assessed in PISA 2003 and PISA 2006 was not aligned with that assessed by the Main NAEP, but that in 2009 the alignment was improved.

FIGURE G

Cumulative gains in NAEP, TIMSS, and PISA mathematics scores, 2000–2011 (standard deviations)



Note: The data points for the LTT NAEP and TIMSS in 2008/2009 are constructed by linear interpolation.

Source: Authors' analysis of National Center for Education Statistics' NAEP Data Explorer; Trends in International Mathematics and Science Study (TIMSS) (Mullis et al. 2001, Mullis et al. 2004, Mullis et al. 2008, Mullis et al. 2012); and OECD Program for International Student Assessment (PISA) (2001, 2004, 2007, 2010a)

We noted above that some U.S. states asked for their TIMSS sample sizes to be increased in some years to generate state-level results. In seven cases data are available on TIMSS scores for states that participated in TIMSS in more than one year, and these permit a comparison of state-level trends on the TIMSS and the Main NAEP, for which there are also state-level trend data.

Table 23 compares trends on Main NAEP for 8th-graders with those on TIMSS in the five cases for which available data permit such comparisons.

The table shows that TIMSS and Main NAEP, at least for these seven states, do not necessarily exhibit similar trends in math. Connecticut, Indiana, Massachusetts, and Minnesota had improving TIMSS and Main NAEP scores during the period for which we can make comparisons. In each of these cases, the rates of improvement on the TIMSS and Main NAEP were quite similar. Massachusetts' rate of improvement was rapid, but believable because each test confirms the other. Minnesota's rate was also rapid (though not as rapid as Massachusetts') and also similar on each test. Connecticut's and Indiana's rates of improvement were very small, but also similar on each test.

Missouri had falling scores on the TIMSS, but its Main NAEP performance was unchanged during the period for which data are available. Oregon had falling scores on the TIMSS, but its Main NAEP performance improved during the period for which data are available.[31] These two states, however, have not participated in TIMSS since 1999, so it would be difficult to explore the underlying causes of these discrepancies.

TABLE 23

Comparing U.S. state-level math trends on TIMSS and Main NAEP, 8th-graders

	1995/1996	1999/2000	2011	Average annual change
Connecticut				
TIMSS		512	518	+0.1%
Main NAEP		281	287	+0.2%
Massachusetts				
TIMSS		513	561	+0.7%
Main NAEP*		279	299	+0.6%
Minnesota				
TIMSS	518		545	+0.3%
Main NAEP	284		295	+0.3%
North Carolina				
TIMSS		495	537	+0.7%
Main NAEP		276	286	+0.3%
Indiana				
TIMSS		515	522	+0.1%
Main NAEP		281	285	+0.1%
Missouri				
TIMSS	505	490		-0.8%
Main NAEP	273	274		0.0%
Oregon				
TIMSS	525	514		-0.5%
Main NAEP	276	281		+0.4%

* The Minnesota 1996 Main NAEP test was administered without accommodations permitted, and the 2011 test was administered with accommodations permitted, so these test results are not strictly comparable.

Source: Harmon et al. (1997); Mullis et al. (1998); Mullis et al. (2001); Mullis et al. (2012); National Center for Education Statistics' NAEP Data Explorer

North Carolina is worth further attention. The state had a very rapid rate of improvement on TIMSS. Among states for which we have data, North Carolina's annual improvement rate of 0.68 percent was second only to that of Massachusetts (0.75 percent) on the TIMSS. But whereas Massachusetts showed similar rates of improvement on the TIMSS and Main NAEP, North Carolina's Main NAEP rate of improvement was less than half its TIMSS rate during the same period.

When Secretary Duncan announced the TIMSS 2011 overall average results in December 2012, he highlighted North Carolina as proving that demographically diverse states can outperform others. Certainly, North Carolina's annual improvement rate on the Main NAEP is impressive. But further study is needed before concluding that the even more impressive TIMSS rate should be believed. As we have pointed out often in this report, the discrepancy between TIMSS and Main NAEP rates of improvement, both of which might nonetheless be substantial, could be the result of a unique curriculum (in this case North Carolina's) more closely aligned with TIMSS than with NAEP, to flaws in sampling either of TIMSS or NAEP, or to other causes.

Table 24 summarizes what we have learned about U.S. students from our examination of the LTT, Main NAEP,

TIMSS, and PISA tests of 13-year-olds, 8th-graders, and 15-year-olds, in reading and mathematics.

For each test, and for each year of available data since 1992, the table shows the average U.S. score and the score for more-disadvantaged students. In the case of the NAEP tests, these are African American students and students whose mothers did not graduate from high school.[32] For PISA and TIMSS, these are students in social class Groups 1 and 2. The next to last right-hand column displays the average annual rate of change in scores for the full period shown. The right-hand column displays the average annual rate of change in scores from the year closest to 2000 for which data are available to the year closest to 2009 for which data are available. Overall, it seems that these tests provide consistent confirmation that U.S. performance has improved more for disadvantaged students than overall, especially in the last decade.

TABLE 24

Comparing U.S. trends for all and for disadvantaged students, Long-Term Trend and Main NAEP, PISA, and TIMSS

	1992	1994	1995	1996	1998	1999	2000	2002	2003	2004	2005	2006	2007	2008	2009	2011	Average annual change, earliest to latest year shown	Average annual change, 1998–2000 to 2007–2009
Reading																		
LTT																		
All	260	258		258		259				258				260			0.0%	0.0%
Mother < HS	242	245		243		241				242				243			0.0	0.1
Af Am	238	234		234		238				241				247			0.2	0.4
Main																		
All	260	260			263			264	263		262		263		264	265	0.1	0.0
Mother < HS	245	244			248			251	249		247		248		250	250	0.1	0.1
Af Am	237	236			243			245	244		243		245		246	249	0.3	0.1
PISA																		
All							504		495						500		-0.1	-0.1
Group 1							418		422						442		0.6	0.7
Group 2							455		457						471		0.4	0.4
Math																		
LTT																		
All	273	274		274		276				280				281			0.2%	0.2%
Mother < HS	259			260		261				265				270			0.3	0.4
Af Am	250	252		252		251				259				262			0.3	0.5
Main																		
All	268			271			274		278		279		281		283	284	0.3	0.3
Mother < HS	252			255			256		260		262		265		267	268	0.3	0.5
Af Am	237			241			245		252		255		260		261	262	0.6	0.7
PISA																		
All							493		483			474			487		-0.1	-0.1
Group 1							416		420			417			434		0.5	0.5

TABLE 24 (CONTINUED)

	1992	1994	1995	1996	1998	1999	2000	2002	2003	2004	2005	2006	2007	2008	2009	2011	Average annual change, earliest to latest year shown	Average annual change, 1998–2000 to 2007–2009
Group 2							446		445			436			464		0.4	0.4
TIMSS																		
All			492			502			504				508			509	0.2	0.2
Group 1						439			449				461					0.6
Group 2						461			473				482					0.6

Note: The TIMSS "all" figure for 1995 comes not from the TIMSS database but is a corrected number as reported in the TIMSS 1999 report. For years when NAEP introduced testing with accommodations, scores shown are averages of results with and without accommodations. Mother < HS is mother's educational status is less than high school, and Af Am is African American.

Source: Authors' analysis of Trends in International Mathematics and Science Study (TIMSS) databases, various years (Boston College International Study Center); OECD Program for International Student Assessment (PISA) databases, various years; National Center for Education Statistics' NAEP Data Explorer

Part VII. Population and curricular sampling issues

In Parts V and VI, we described serious inconsistencies in the achievement trends for U.S. disadvantaged and advantaged students on several international and national tests. Such conflicting results suggest caution about drawing policy inferences without delving more deeply into what these tests measure. But beyond conflicting results among various evaluations of student learning, each test has its sampling peculiarities that can affect results. Some of these sampling peculiarities, such as the oversampling of U.S. disadvantaged students in high-poverty high schools in PISA, can bias the results to a degree that we can estimate. Other aspects of the tests, such as the greater tendency of students in some countries to random mark rather than leave answers blank, can also bias results in ways that we cannot estimate.

In most cases, it is not possible to re-estimate U.S. scores to account for elimination of such problems. But we can adjust for the effect on scores of the unusually disadvantaged sample of U.S. test takers and of a compounding of this effect by an oversampling of the most disadvantaged U.S. students in the PISA sample. We conclude that correcting for these two problems would improve the U.S. average score and international rank in both reading and mathematics; in the case of mathematics it improves the average score substantially.

Test makers also make decisions about how to sample the curriculum, and these decisions affect how countries' performances compare. For example, if one country's students do better in algebra than geometry, and another's do better in geometry than algebra, the first country will appear to have better math performance on a test that has a higher proportion of algebra questions and worse on a test that has a higher proportion of geometry questions. We have limited ability to make precise adjustments of international (or interstate) comparisons for these decisions, but we can show that they affect common judgments about relative national performance.

In this section, we review these various conflicts, flaws, and other possible biases in test results that suggest the need for caution in interpreting average national test score differences as valid measures of the comparative quality of U.S. schools.

Population sampling flaws

None of the assessments to which we refer in this report, PISA, TIMSS, LTT, or Main NAEP, are universally administered to all students of the appropriate age or grade level in a country. Rather, the test is given to a small sample, but one that statisticians deem large enough to be representative of all students. The larger the sample, the more representative it can be. PISA, for example, constructed samples that were large enough for analysts to be confident of a 95 percent probability that results in the United States for reading are within about 7.5 points (two standard errors), and in mathematics about 7 points (two standard errors), of results that would be obtained if the test were given to all students.[33]

For each PISA test administration, it is necessary for each nation to determine a necessary sample size and then make a random selection of its 15-year-olds. If the sampling process is flawed, the reported results can be quite inaccurate. For example, if the proportion of low achievers in a country who take the test is higher than the proportion of low achievers in the nation as a whole, the reported "average" score will be artificially low, and not truly representative of that country's performance.

The sampling methodology is complex, and the possibility of sampling flaws is another reason why results should be treated with caution. Sampling requires selecting schools that are large enough to have a sufficient number of 15-year-olds and that seem to be representative of geographic regions; public and private schools; rural, suburban, and urban schools; schools with minority populations; and a few other characteristics.

Unfortunately, in 2009 a sampling flaw in the United States seems to have produced a PISA sample whose aver-

age score was lower than the average score would have been from an accurately representative U.S. sample.

PISA reports that 35 percent of its test takers were eligible for the free and reduced-price lunch (FRPL) program. The National Center for Education Statistics reports that 38 percent of all U.S. high school students were FRPL eligible during the 2009–2010 school year in which the PISA test was administered (NCES online). In this respect, the sample seems representative.

However, it is not sufficient to have a representative proportion of FRPL-eligible students in the overall sample, because we know that disadvantaged students perform more poorly if they attend schools where they are not integrated with more advantaged students and are instead heavily concentrated with other FRPL-eligible students. Controlling for a student's own family income, those who attend high-poverty schools are less likely to benefit from positive modeling of higher-achieving peers, are more likely to suffer from the stress of violent neighborhoods, are more likely to experience disruptions where instructional resources are diverted to discipline, are more likely to lose continuity of instruction when teachers repeat lessons for the benefit of more mobile newcomers, are less likely to benefit from school and instructional policies monitored by more involved parents, and are more likely to have less experienced teachers. These characteristics of high-poverty schools frequently result in lower achievement for students who attend such schools.[34]

Students who attend schools where disadvantage is concentrated are likely to perform, on average, at considerably lower levels than students whose family income is similarly low but who attend schools where more students are middle class. A sampled population that includes students eligible for FRPL who are dispersed across many schools will typically have higher average achievement than a similar sampled population with the same proportion of FRPL students but where these students are concentrated in fewer schools.

Therefore, for an accurate sample, PISA should not only have a proportion of FRPL-eligible students that is similar to that proportion nationwide, but should have FRPL-eligible students whose distribution among schools with concentrated disadvantage is also similar to the distribution nationwide.

Table 25 compares the distribution of all U.S. high school students nationwide, by share of FRPL-eligible students in their high schools, to the distribution of students in the 2009 PISA sample, by share of FRPL-eligible students in their high schools.[35]

The table shows that the average PISA score of U.S. students in both reading and math decreases dramatically as the share of their schools' students who are FRPL-eligible increases. The table also makes apparent that PISA's FRPL test takers were heavily concentrated in severely disadvantaged schools, where unusually large proportions of students were FRPL-eligible. Forty percent of the PISA sample was drawn from schools where half or more of the students were eligible for free or reduced-price lunches. Only 32 percent of all U.S. students attended such schools in 2009–2010 when the PISA test was given. Sixteen percent of the PISA sample was drawn from schools where more than 75 percent of students are FRPL-eligible, yet fewer than half as many, 6 percent of U.S. high school students, actually attend schools that are so seriously impacted by concentrated poverty.

Likewise, students who attend schools where few students are FRPL-eligible, and whose scores tend, on average, to be higher, were undersampled. This oversampling of students who attend schools with high levels of poverty and undersampling of students from schools with less poverty results in artificially low PISA reports of national average scores.

If other countries' PISA samples better reflect the actual spatial distribution of disadvantaged 15-year-olds, the real U.S. average performance should rank higher relative to other countries than the reported PISA averages indicate.

TABLE 25

Shares of all U.S. high school students, and of PISA sample, by free and reduced-price lunch (FRPL) percentages of their schools, with average PISA reading and math scores, by FRPL percentages of schools

	(a)	(b)	(c)	(d)
Share of students eligible for FRPL in student's school	**Share of all U.S. high school students, by share of students in school who are FRPL-eligible, 2007–2008**	**Share of PISA 2009 sample in high schools, by school percent of students eligible for FRPL**	**Average U.S. PISA reading score, by school percent of students eligible for FRPL**	**Average U.S. PISA math score, by school percent of students eligible for FRPL**
(a) 75 percent or more	9%	16%	449	437
(b) 50 to 74.9 percent	22	24	472	461
(c) 25 to 49.9 percent	36	36	502	488
(d) Less than 25 percent	30	24	538	533
(e) No data available	3			
(f) All	100%	100%	495	484
(g) PISA average scores, weighted by actual share of schools with specific FRPL-eligible ranges (columns c and d, weighted by column a)			501	491

Source: NCES (2012), Table A-13-1 (for column a); authors' analysis of OECD Program for International Student Assessment (PISA) 2009 database (for columns b–d)

We have queried officials at the National Center for Education Statistics in an attempt to determine why the PISA sample was skewed in this way, but while these officials acknowledge that there may be a sampling error, they have been unable to provide an explanation.[36] We can only speculate about it. One possibility is that the PISA sampling methodology excluded very small schools, where poverty is less likely to be concentrated. Another possibility is that because participation in PISA is voluntary on the part of schools and districts that are randomly selected for the sample, schools serving more affluent students may be more likely to decline to participate after being selected. Perhaps this is because such schools are generally less supervised by the federal government than schools serving disadvantaged students and feel freer to decline government requests. Whatever the reason, an initial PISA sample that was representative would lose some validity if schools serving higher proportions of more affluent children were more likely to decline to cooperate, and were then replaced in the sample by schools serving lower proportions of affluent students. An underestimation of national average scores is then bound to result.

To get a sense of how much of an underestimate resulted, we recalculated the overall U.S. average reading and math PISA scores, using the data in Table 25. For this recalculation, we assume that the average score of students attending schools in each category of FRPL participation is unchanged, but the proportion of such students is that of the nation as a whole, not that of the PISA sample. We find that with these assumptions, the U.S. reading and math scores would be about the same (1 scale point higher, or 501 rather than 500 in reading; and 3 scale points higher, or 491 rather than 487 in math).

Indeed, the effect of the sampling error is probably even greater, because 3 percent of schools nationwide do not report their FRPL percentages to the National Center for Education Statistics. It is more likely that these schools are those without any FRPL-eligible students, because schools that do not participate in government programs are more likely to fail to comply with reporting requirements. If so, the missing data probably come from schools whose average scores are somewhat higher than those from schools that did report but that had few FRPL-eligible students (538 in reading and 533 in math, from

row (d) of Table 25). Then, the calculations in row (g) of Table 25 would yield averages that are higher than 501 and 491.

In Part II, we showed (see Table 2A) that the U.S. sample was more heavily weighted toward disadvantaged students (Groups 1 and 2) and more heavily weighted against advantaged students (Groups 5 and 6) than the samples of our six comparison countries. Then we showed, for example (see Tables 3B and 3D), that if the social class distribution of the U.S. sample was similar to the average social class distribution of the three similar post-industrial comparison countries, the average U.S. reading score would have been better, 9 scale points higher (jumping from the reported average U.S. reading score of 500 to the social-class-adjusted U.S. score of 509), and the average U.S. math score would also have been better, 8 scale points higher (jumping from 487 to 495).

If we add the two social class adjustments together, one for the excess preponderance of disadvantaged students in the U.S. sample (in comparison to similar post-industrial countries), and one for the oversampling of students from schools with concentrated disadvantage, we can conclude that a more accurate and comparable average U.S. PISA reading score might have been better, 510 (500 + 1 point + 9 points), and a more accurate and comparable U.S. PISA math score might also have been better, 499 (487 + 8 points + 3 points).

As noted above, these adjusted average scores may still be too low, because if disadvantaged students had been sampled accurately in schools with less concentrated disadvantage, the average scores of U.S. disadvantaged students would likely be somewhat higher. But this consideration is offset by another: When we adjust the U.S. scores for the lower proportion of disadvantaged students in comparison countries, we implicitly reduce the proportion of disadvantaged students in the U.S. population. When the proportion of disadvantaged students decreases, the potential for bias in the average test score from oversampling in high-poverty schools also decreases, simply because the weight of disadvantaged students in the average national score is lower. Thus, the adjustment we make for sampling error, and the adjustment we make for the proportion of disadvantaged students in the total sample, will overlap, but we cannot say to what extent. On balance, taking these two considerations together, we consider the adjusted reading and math scores of 510 and 499 to be plausible.

As a not unreasonable speculative exercise, if we accept this adjusted average U.S. reading score (510), we would conclude that U.S. average PISA reading performance in 2009 was higher than the average performance in each of the similar post-industrial countries, but still not as high as the average performance in Canada and substantially below the average performance in Finland and Korea. If we accept this adjusted average U.S. math score (499), we would conclude that U.S. average PISA math performance in 2009 was about the same as the average math performance in the similar post-industrial countries of France, Germany, and the United Kingdom, but still substantially below the average math performance of the top-scoring countries.

In this report, we have focused only on the United States and six comparison countries. However, in discussions of PISA scores, the media and policymakers have frequently cited the fact that of all 34 OECD test-taking countries in 2009, the United States ranked 14th in reading and 25th in math. If we use our adjusted (for social class composition and sampling error) U.S. scores of 510 for reading and 499 for math, and assume that average scores with scale point differences of less than 8 are about the same (even where OECD reports them as "statistically significant"), we find that the United States would have ranked sixth internationally in reading and 13th in math.

In reading, only Canada, Finland, Japan, Korea, and New Zealand had scores higher than 510. In math, Australia, Belgium, Canada, Estonia, Finland, Germany, Iceland, Japan, Korea, the Netherlands, New Zealand, and

TABLE 26

Comparing U.S. social class changes in PISA and TIMSS, 1999 (2000) to 2007

	PERCENTAGE OF TEST TAKERS			PERCENTAGE OF TEST TAKERS		
	PISA			TIMSS		
	2000	"2007"	Percentage-point change	1999	2007	Percentage-point change
Group 1 (Lowest)	13%	17%	4	8%	17%	9
Group 2	13	16	3	14	20	6
Group 3	27	28	1	29	28	0
Group 4	20	18	-3	22	17	-5
Group 5/6 (Highest)	26	20	-6	28	18	-10
Disadvantaged (Groups 1 and 2)	27	34	7	22	37	15

Source: Authors' analysis of Trends in International Mathematics and Science Study (TIMSS), 1999 and 2007 databases (Boston College International Study Center) and OECD Program for International Student Assessment (PISA) 2000, 2006, and 2009 databases

Switzerland had average scores higher than the adjusted U.S. score of 499.[37]

Population sampling inconsistency between tests

Employing sophisticated sampling techniques, the IEA (for TIMSS) and the OECD (for PISA) both base their results on what they consider accurate samples of national populations.

Yet unless our claim is seriously flawed that BH is the most reasonable proxy available for social class characteristics relevant to student academic performance, it is apparent that either TIMSS or PISA, or both, have failed to administer tests to accurate samples of national populations and that, therefore, the national average score results reported in TIMSS or PISA, or both, should not be taken as accurate. We have already noted that differences in the social class composition of PISA 2000 test takers in reading vs. math themselves cast doubt on the accuracy of reported average results.

Table 2A showed the distribution of 2009 test takers in the United States (and other countries) by social class group. Table 8A showed how the distribution by social class in the United States (and other countries) changed from 2000 to 2009.

Table 26 repeats similar data for PISA and TIMSS, with some changes.

For purposes of comparison with TIMSS, not given in reading, Table 26 uses, for 2000, the books-in-the-home distribution for the PISA math test only. And instead of reporting PISA social class distribution for 2009, it estimates social class composition for PISA "2007" (by averaging social class compositions for PISA 2006 and PISA 2009, with PISA 2006 given twice the weight of PISA 2009). And third, for ease of comparison with TIMSS, whose highest BH category combines the two highest categories in PISA, PISA social class Groups 5 and 6 have been combined. Table 26 then calculates the change in social class composition for PISA mathematics test takers from 2000 to 2007, and adds similar data for TIMSS from 1999 to 2007, almost an identical period.

If books in the home is a reasonable proxy for social class characteristics relevant to student academic performance, then there are apparently flaws in the student samples to which either the TIMSS or PISA, or both, were administered. According to PISA sampling, the share of students

who were disadvantaged (Groups 1 and 2) increased from 2000 to 2007 by 27 percent. According to TIMSS sampling, the share of the same students over almost the same time period increased by 70 percent. There are also important differences between changes from 2000 to 2007 in the relative sizes of social class Groups 4 and 5/6 in the PISA and TIMSS samples.

It is important to remember that these sampling inconsistencies do not call into question the accuracy of the average scores for each of the social class groups in either TIMSS or PISA. They do, however, call into serious question the accuracy of the national average reported scores, and it is these scores to which policymakers and pundits direct such anguished attention.

As noted, NAEP does not report books-in-the-home data for test takers. It does report the share of the sample that participated in the free and reduced-price lunch program and the share of the sample where the test taker's mother had less than a high school education. These are not comparable to BH, but do indicate something about test takers' social class composition. From 2000 to 2007, the share of 8th-grade Main NAEP math test takers whose mothers had only a high school education or less declined from 34 to 30 percent (i.e., mothers' educational backgrounds improved). From 2000 to 2007, the share of 8th-grade Main NAEP math test takers who participated in the free and reduced-price lunch program increased from 29 to 37 percent. We tend to think that the educational attainment of mothers is a more relevant (for comparison with BH data) factor than receipt of free or reduced-price lunches, so we suspect that the NAEP data are more consistent with PISA data that show the share of disadvantaged students increasing by 27 percent in this period than with TIMSS data that show this share increasing by 70 percent. If so, then the TIMSS reported average score may have been erroneously low in 2007, because TIMSS sampled too many lower-scoring disadvantaged students. Or, perhaps, the score was erroneously high in 1999, because it sampled too few lower-scoring disadvantaged students.

Either of these errors would have dampened the report of a real increase in TIMSS scores from 1999 to 2007. Recall that Figure G showed that the Main NAEP U.S. average scores increased at a more rapid rate in the period than did U.S. TIMSS scores. If we are correct about this possible error in TIMSS sampling, then the slope of the TIMSS scores from 1999 to 2007 would have been steeper and more similar to that in the Main NAEP.

Decisions about curricular sampling

From what we have seen so far, it is apparent that no single assessment accurately reflects student performance. In Part V, we compared results from PISA with those from TIMSS for the nations on which we focus in this report, as well as for U.S. states, Canadian provinces, and U.K. countries for which we have data. A look at anomalies from the scores of many other countries, not examined in detail in this report, provides further evidence. Students in Australia, Slovenia, and Norway performed better or substantially better than U.S. students on the PISA mathematics test in both 2006 and 2009, but performed worse than or only as well as U.S. students on the 2007 and 2011 TIMSS mathematics tests. Students in New Zealand performed substantially better than U.S. students on the PISA mathematics test in both 2006 and 2009, but performed substantially worse than U.S. students on the TIMSS 2011 test. (New Zealand did not participate in the 2007 TIMSS administration.) Eleven other countries that performed better than the United States on the PISA mathematics test in 2009 did not participate in TIMSS 2007 or 2011, so we cannot know how widespread inconsistent relative performance on PISA and TIMSS might have been if all countries participated in both tests.

Other inconsistencies appear when we compare trends in U.S. scores on PISA with trends in scores on the NAEP. As Figure G illustrates, U.S. average math scores plummeted on PISA from 2000 to 2003, and then

plummeted further from 2003 to 2006. But U.S. PISA mathematics scores made rapid gains from 2006 to 2009, so the U.S. average PISA math score in 2009 was almost back to its 2000 level. But on the Main NAEP, 8th-grade math scores increased consistently from 2000 to 2009, with the rate of increase more rapid in the first half of the decade, the very years when U.S. PISA math scores were falling. From 2000 to 2006, TIMSS math scores remained about the same while U.S. PISA math scores were falling substantially and U.S. Main NAEP math scores were rising substantially.

As discussed above, these inconsistencies could result from flaws in population sampling. If a test samples a larger proportion of disadvantaged students than is present in the national student population, it could erroneously report national average performance that is lower than another test with a more accurate sample. Yet even if population samples were accurate in both tests, they could report inconsistent average performance because of a different kind of sampling problem—inconsistencies in curriculum sampling, i.e., what topics or skills in math and reading the tests emphasize. No test of an hour or so can assess every topic or skill in the curriculum; test designers must make judgments about which topics or skills to include, and what emphasis each should be given. A possible explanation for the inconsistencies between the tests discussed in this report could be that each assessment samples different aspects of the mathematics or reading curriculum. PISA, for example, has relatively more problem-solving than computational items, compared to TIMSS and NAEP. If results on these tests seem inconsistent, it may be because one is better aligned with a country's curriculum than the others, or because a country's teachers are relatively more effective with some parts of the curriculum than with others.

Because PISA includes a larger proportion of more practical problem-solving items, relative to items requiring only computation, many experts consider the PISA test to be a better and more sophisticated mathematics test than other standardized tests like TIMSS or NAEP.

Yet the actual data from these tests illustrated in Figure G challenge this conventional description of curricular test differences. Because the Main NAEP includes more problem-solving and constructed response items than the LTT (which has more stress on basic computation), we might expect U.S. trends on the Main NAEP to be more similar to U.S. trends on the PISA than to U.S. trends on TIMSS. But they are not, or at least not consistently. Indeed, U.S. trends on the Main NAEP and LTT mathematics tests are very similar (especially in comparison to trends on TIMSS and PISA), suggesting either that the U.S. curriculum is exquisitely balanced between problem solving and computation, or that the differences in curricular coverage between the Main NAEP and LTT are not very great while the differences in curricular coverage are great between both NAEP assessments and the PISA and somewhat less great between both NAEP assessments and the TIMSS.

Commonplace explanations of why tests can differ so much in their reports of student performance are not persuasive. For example, some U.S. education experts believe that with PISA having more emphasis on application of math to "real world" problem solving, TIMSS is more closely aligned with the U.S. math curriculum than is PISA (Robelen 2012a, 2012b). But this does not seem to be the case. As Table 27 shows, the share of the TIMSS 8th-grade test devoted to geometry increased by two-thirds from 1999 to 2011, while the share devoted to algebra increased only by one-third. Yet few American students study geometry intensively in 8th grade (it is typically given greater attention in the 10th grade), while there have been efforts across the United States to ensure that all students are introduced to algebra in the 8th grade. The Main NAEP is specifically intended to reflect the U.S. math curriculum. So claims that discrepancies between U.S. results on PISA and TIMSS can be attrib-

utable to TIMSS being more aligned than PISA with the U.S. curriculum require a stronger foundation.

Further, if the allegations of PISA sophistication are correct, this sophistication may have a downside. Because of the large number of problem-solving items, the PISA math assessment is effectively a reading comprehension test as well as a mathematics test. Because parental literacy has a big impact on children's reading ability, social class differences may have a larger impact on differences in reading ability than on differences in mathematics proficiency. If so, PISA may more accurately reflect how well the math curriculum was delivered to upper- than to lower-class students. Alternatively, countries with more effective literacy instruction may have an advantage on PISA's mathematics assessment, independent of the quality of math instruction.

None of these considerations, however, help to explain the curious V-shape of the PISA results in Figure G. If consistent differences in curricular alignment between tests were the causes of different test trends, we would not expect PISA results to diverge sharply from the Main NAEP results from 2000 to 2006 and then to parallel those results from 2006 to 2009.

Scholars have not explained the apparent trend inconsistencies between PISA, TIMSS, and NAEP, nor have they considered whether these inconsistencies threaten the validity of either the PISA or TIMSS test for other countries.[38]

Consider the choices made by TIMSS test designers in the topics to sample. Before tests are developed, test sponsors (in this case the IEA) develop instructions for test developers regarding what topics should be covered and in what proportion. The actual tests generally adhere to these instructions. For 1999, we have information on the actual proportion of items in the various content areas that appeared on the TIMSS. For subsequent years, we have information only on the target instructions for the test developers. We have no reason to believe that there are important differences between the target and actual proportions.

Table 27 shows the proportion of TIMSS 8th-grade mathematics tests devoted to different aspects of mathematics content.

TABLE 27

Content coverage in the 8th-grade mathematics assessment, TIMSS 1999, 2003, 2007, and 2011 (percentages devoted to each topic)

	1999*	2003**	2007**	2011**
Numbers	37%	30%	30%	30%
Data representation	13	15		
Data and chance			20	20
Geometry	12	15	20	20
Algebra	22	25	30	30
Measurement	15	15		

* Post assessment analysis of actual items used in assessment

** Target instructions to test developers

Source: Martin and Mullis (2001); Mullis et al. (2003); Mullis et al. (2005); and Mullis et al. (2009)

We can see that the content categories differed in 2007 and 2011 from those used in 1999 and 2003. We cannot judge whether any of these categories are exactly comparable between the first two and second two administrations, or how item types in the 1999–2003 categories were redistributed into the 2007–2011 categories, but we think it likely that the importance of numbers (e.g., fractions, decimals) was reduced, that measurement (e.g., perimeter, area, volume) was partially reduced and partly shifted to geometry; that the importance of algebra was increased, and that the importance of probability and statistics (what TIMSS terms "data and chance") was increased, with some of this increase attributable to the redistribution of some data representation items to the data and chance category. If these are the case, then countries that place more emphasis on probability and statistics, algebra, and geometry, or whose students do better in these areas, will have the opportunity to record greater apparent growth on TIMSS over time than countries that

place more emphasis on numbers and simple measurement.

Table 28A shows how U.S. students performed on these distinct content areas over time.

TABLE 28A

Average 8th-grade scale scores, U.S., by mathematics content area, TIMSS 1999, 2003, 2007, and 2011

	1999	2003	2007	2011
Numbers	509	508	514	514
Data representation	506	527		
Data and chance			533	527
Geometry	473	472	480	485
Algebra	506	510	507	512
Measurement	482	495		

Source: Mullis et al. (2000); Mullis et al. (2004); Mullis et al. (2008); Mullis et al. (2012)

TABLE 28B

Average 8th-grade scale scores, Finland, by mathematics content area, TIMSS 1999 and 2011

	1999	2011
Numbers	531	527
Data representation	525	
Data and chance		542
Geometry	494	502
Algebra	498	492
Measurement	521	

Source: Mullis et al. (2000); Mullis et al. (2012)

According to the table, U.S. 8th-graders' strongest area is probability and statistics, a topic that has probably increased in importance since 2003. They also do relatively well in algebra. But U.S. students do relatively poorly in geometry, a topic that has also increased in importance. If, for example, the weight of probability and statistics had been further increased at the expense of the increase in geometry, U.S. average scores would have improved in this period without any improvement in the quality of instruction. But it is also the case that if teachers spent more time on algebra and statistics, and less on geometry, their efforts would be rewarded beyond the additional learning taking place, simply because the weights in the test had changed.

Table 28B displays similar performance data for Finland, which participated in TIMSS only in 1999 and 2011.

We noted above that Finland's average 8th-grade math TIMSS performance had fallen so that it is now about the same as that of the United States. Table 28B suggests that this may be attributable to Finland's failing to update its curriculum in line with greater contemporary emphasis on algebra and geometry. Finland does relatively less well on these topics than on simple number properties. Finland does do relatively well, as does the United States, on statistics and probability, but this may apparently not be sufficient to offset the greater emphasis now given on TIMSS to algebra and geometry. If, however, the weights in the TIMSS had not changed, it is possible that Finland would still appear to perform better, on average, than the United States.

Comparing U.S. national performance with that of Finland, policymakers will be surprised to learn that U.S. students now substantially outperform Finnish students in algebra.

A puzzling consequence of this interpretation, however, is that the average performance of Finland on the PISA mathematics test is much superior to that of the United States. Inasmuch as PISA is reputed to be a more challenging test than TIMSS, it is curious that Finland would perform relatively better (than the United States) on PISA when its performance in algebra is relatively worse than its performance on simple number property problems. Further investigation of this incongruity is certainly in order before definitive conclusions can be reached about the relative performance of the two countries.

Finally, **Table 28C** examines the relative 8th-grade mathematics performance by content areas of all the U.S. states that participated in the 2011 TIMSS.

TABLE 28C

8th-grade mathematics performance for U.S. and selected U.S. states by mathematics content area, TIMSS 2011

	Alabama	California	Colorado	Connecticut	Florida	Indiana	Massachusetts	Minnesota	North Carolina	U.S.
Numbers	463	492	521	527	517	528	567	556	547	514
Data and chance	480	495	540	546	528	545	584	571	548	527
Geometry	443	454	505	490	499	498	548	515	515	485
Algebra	471	509	512	510	513	520	559	543	537	512
Average score	466	493	518	518	513	522	561	545	537	509

Source: Mullis et al. (2012)

Table 28C seems to show a remarkable consistency in the curricula of these selected states, despite frequent complaints by policymakers that the United States is disadvantaged by the lack of a national curriculum. (Whether this conclusion would be sustained if all states participated in TIMSS is unknown.) In each state shown, as in the United States nationwide, students perform less well, compared to students internationally, in geometry. Almost as consistently, students in each state and in the United States nationwide perform relatively best, compared to students internationally, in probability and statistics. The only exception is California, where students perform relatively best in algebra, with statistics next.

This should be encouraging to those policymakers who have advocated increased emphasis in U.S. schools on probability and statistics, in part because this skill is essential to good citizenship. Students also do relatively well, compared to students internationally, in algebra, perhaps indicating that intense policy advocacy on introducing algebra in 8th grade has had an effect.

It would be tempting to think that the United States could increase its international standing in mathematics by encouraging educators to pay more attention to geometry. Likewise, the United States could increase its international ranking by advocating, within the IEA, for giving greater weight in the next TIMSS to statistics and less to numbers. This would not necessarily be wise, however. Education policymakers should make choices about curricular priorities based on what is best for the nation, not on what can generate artificial gains on internationally comparative tests. But they should also keep in mind that the relative weights displayed in Table 27 are the result of policy judgments that reflect a consensus of experts from many countries and are not necessarily those that the United States would choose were it solely responsible. To the extent that this international policy consensus differs from a U.S. policy decision, relative scores on an international test like TIMSS tell us less than we usually think about how U.S. students perform relative to those in other countries.

Assessment by age or by grade

Another complexity is that PISA is administered to a representative sample of 15-year-olds, regardless of their grade in school. But TIMSS is administered to a representative sample of 8th-graders, regardless of their age. Because not all countries enroll students in kindergarten or the first grade at the same age, 15-year-olds in some countries have had more schooling than in others. Some countries may also have more severe retention policies than others, resulting in a larger proportion of 15-year-olds in earlier grades. As a result of both factors—international differences in school entry ages and retention policies—interpretation of both PISA and TIMSS results must disentangle the effects of a country's grade progression policies from the effectiveness of its

TABLE 29

Share of sample in each grade, for U.S. and six comparison countries, PISA 2009

	Canada	Finland	Korea	France	Germany	U.K.	U.S.
7th grade	0%	1%	0%	1%	1%	0%	0%
8th grade	1	12	0	4	11	0	0
9th grade	14	87	4	34	55	0	11
10th grade	84	0	95	57	33	1	69
11th grade	1	0	1	4	0	98	20
12th grade	0	0	0	0	0	1	0

Source: OECD Program for International Student Assessment (PISA) (2010a), Table A2.4a, p. 180

mathematics (and, in the case of PISA, reading) instruction.

Countries vary greatly in the percentage of 15-year-olds sampled for PISA in various grades. **Table 29** displays the grade distributions of PISA test takers in the United States and comparison countries.

In the United States, 69 percent of the 15-year-olds tested in the 2009 PISA administration were 10th-graders, 20 percent were 11th-graders and 11 percent were 9th-graders. Because students start kindergarten later in Finland and few are retained, almost all tested 15-year-olds (87 percent) in that country were in 9th grade. In Korea and Canada, almost all were in 10th grade, and, in the United Kingdom, virtually all were in the 11th grade. In France and Germany, the sample was more equally spread between the 9th and 10th grades.

Grade level may make a difference in how well students perform. **Table 30** shows the differences, for the United States and each comparison country, in average PISA reading and math scores of students at the various grade levels.

In each country, 15-year-olds in higher grades perform better than 15-year-olds in lower grades. The modal grade for each country is bolded and underlined.

Note especially that Finland, an unusually high-scoring country, enrolls most 15-year-olds in an earlier grade than do most other countries. So at first glance, it may seem that Finland's scores are especially noteworthy because its students have higher scores despite having been in school a shorter length of time. However, almost all Finnish children attend publicly subsidized early childhood programs from the age of 2. Thus, by the time they enter 9th grade, they have been in organized school environments for 13 years, longer than children in comparison countries who did not benefit from similar early childhood education. Thus, if anything is to be learned in this respect from Finland's high scores, it is not that lower-scoring countries start schooling too early, but rather that they start it too late.

The other country with an unusually high proportion of 15-year-olds in the 9th grade rather than the 10th is Germany. As we have seen, although Germany has achieved much more rapid improvement in achievement since 2000 than any other country we have studied, its scores remain relatively low. Perhaps this low level is simply the result of a national educational system that chooses to enroll children at a later age than other national systems. If this were the case, then German results might appear to be relatively superior, internationally, if PISA (like TIMSS, or Main NAEP) had been administered to students at a given grade rather than to students at a given age.

Possibly, in some countries the distribution by grade of PISA test takers may reflect only the social class distribu-

TABLE 30

Reading and mathematics scale scores, by grade, U.S. and six comparison countries, PISA 2009

	Canada	Finland	Korea	France	Germany	U.K.	U.S.
Reading scale score							
8th grade	395	494		380	420		
9th grade	483	**542**	515	417	**489**		414
10th grade	**532**		**540**	**545**	551	452	**506**
11th grade	591			561		**494**	527
12th grade						545	
Mathematics scale score							
8th grade	410	500		381	430		
9th grade	491	**546**	515	423	**502**		410
10th grade	**533**		**548**	**543**	570	451	**493**
11th grade	598			572		**493**	510
12th grade						540	

Note: Bold and underlined numbers show the mean score for the modal grade for the PISA sample in each country.

Source: Authors' analysis of OECD Program for International Student Assessment (PISA) 2009 database for each country

tion, if families of different social classes tend to enroll their children in school at different ages, or if students of different social classes are more or less likely to be retained. We have not, however, performed an analysis to determine the extent to which there is a relationship in each country between students by social class group and students by grade level. Our assumption, however, is that there is some relationship between PISA scores and grade level, independent of the relationship between PISA scores and social class group.

But even if the relationship of PISA scores to grade level were independent of social class group, the different grade distribution of students in different countries may not itself fully explain differences in average performance. The direction of causality in the grade level/performance relationship is unclear. Do students in higher grades perform better because they have been exposed to more instruction? Or do students who perform more poorly get held back into lower grades? If countries permit parents some discretion in the age at which they first enroll children in school, is there a relationship between age-in-grade and social class group attributable to this discretion? Without answers to these questions, we cannot make any definitive statements about the "effect" on PISA scores of differences in students' average grade level among countries.

In addition, because PISA samples students at a common age, not a common grade, the grade distribution of test takers in a country can depend on the month of the year in which PISA is administered. Countries have not always administered successive PISA tests in the same month of the year, producing another challenge to analysts trying to make sense of changes in PISA test scores.[39]

Part VIII. Discussion

As noted in the introduction, Education Secretary Duncan called the 2011 TIMSS results "unacceptable," saying that they "underscore the urgency of accelerating achievement in secondary school and the need to close large and persistent achievement gaps" (Duncan 2012). Two years before he said that the 2009 PISA results "show that American students are poorly prepared to compete in today's knowledge economy. … Americans need to wake up to this educational reality—instead of napping at the wheel while emerging competitors prepare their students

for economic leadership." Referring to the PISA results for disadvantaged U.S. students, Duncan said: "As disturbing as these national trends are for America, enormous achievement gaps among black and Hispanic students portend even more trouble for the United States in the years ahead. Last year, McKinsey & Company released an analysis which concluded that America's failure to close achievement gaps had imposed—and here I quote—'the economic equivalent of a permanent national recession.'" The PISA results, Duncan concluded, justify the reform policies he has been pursuing: "I was struck by the convergence between the practices of high-performing countries and many of the reforms that state and local leaders have pursued in the last two years" (Duncan 2010).

A prominent proponent of the argument that international score comparisons support the need for radical U.S. school reform has been Andres Schleicher, director of the PISA program of the OECD. Duncan consults with and has been influenced by Schleicher in the design of his U.S. education policies. Schleicher asserts that "international education benchmarks make disappointing reading for the U.S." (Dillon 2010) and that "in the U.S. in particular, poverty was destiny. Low-income American students did (and still do) much worse than high-income ones on PISA. But poor kids in Finland and Canada do far better relative to their more privileged peers, *despite* their disadvantages" (Ripley 2011).

We have shown that this claim is untrue. Simple calculations from Table 5, above, show that, on the 2009 PISA reading test, the ratio of average scores of the lowest social class group to the highest social class group in the United States was 0.78, and in Canada and Finland it was 0.81. On the 2009 math test, the ratios were 0.79, 0.83, and 0.85, respectively. These are better ratios in Canada and Finland, but not "far better," and considering the unusually high concentration of disadvantaged students in some U.S. schools, a concentration not found in schools in Canada and Finland, it is surprising that the differences are not greater. Schleicher testified before a U.S. congressional committee considering the reauthorization of the Elementary and Secondary Education Act that Finland had the world's "best performing education system" (Dillon 2010), but he made no reference to the deterioration of Finnish performance for almost all social class groups in reading since 2000, shown above in Table 10A, and for disadvantaged students in math, shown above in Table 12A. At the time Schleicher testified, of course, he could not have known that by 2011, national average 8th-grade mathematics scores in Finland would be about the same as those in the United States and perhaps worse, once differences in the two countries' social class composition was controlled for.

In the wake of the PISA 2009 score release, Secretary Duncan requested that the OECD prepare a report on lessons for the United States from international test data. In response, the OECD advised him that U.S. students have "a significant advantage compared with other industrialised countries" on a range of social class indicators, and, therefore, U.S. students should be expected to perform better than they do. The report argues that "[a] comparison of the percentage of 35-to-44-year-olds that have attained upper secondary or tertiary levels of education, which roughly corresponds to the age group of parents of the 15-year-olds assessed in PISA, ranks the United States 8th among the 34 OECD countries" and that "[t]he share of students from disadvantaged background in the United States is about average" (OECD 2011, 26-28).

Yet in all six of our comparison countries—four of which the OECD report cites as examples from which lessons can be learned for improving U.S. education (Canada, Finland, Germany, and the United Kingdom)—the sample of students taking the PISA test averages much higher levels of books in the home than in the U.S. sample, and thus the share of U.S. students from disadvantaged backgrounds is much higher than in any of the comparison countries. The OECD conclusion that the share of disadvantaged students in the United States is

"about average" results from its reliance on questionable measures to define relative inter-country disadvantage—in particular, material possessions and years of parental schooling. As noted above, relative parental education levels certainly affect educational achievement of students within a country, but differences in parental education levels between countries may not reflect social class differences.

The OECD report, however, then contradicts itself and proceeds to deny the relevance of social class entirely. It states that "…the future economic and social prospects of both individuals and countries depends on the results they actually achieve, not on the performance they might have achieved under different social and economic conditions. That is why the results that are actually achieved by students, schools and countries are the focus of the subsequent analysis in this chapter." For this reason, the OECD report bases its assessment of relative U.S. performance only on countries' average performance, not on that of different social class groups.

Our analysis points policymakers in a very different direction. We argue that policymakers should draw lessons from educational reforms that improve student learning in particular social environments and that show sustained success in such environments over time. Undoubtedly, for example, Finland has been successful in producing high academic performance. Many of the lessons of the Finnish educational system—relatively high teacher salaries, excellent teacher training, the high status of the teaching profession that encourages many highly qualified young people to become teachers—are valuable in understanding how to improve U.S. education. At the same time, the academic performance on PISA of Finnish students has dropped significantly in the last decade, especially for disadvantaged students. This cannot be because Finland was overwhelmed with immigrants having low levels of literacy in 2009, compared with 2000: According to PISA, the share of Finnish students in the disadvantaged social classes (Groups 1 and 2) declined from 27 percent to 17 percent in this period. If accurate, this decline should have made it easier for Finnish educators to concentrate on those disadvantaged students who remained, unless a seemingly large increase in family literacy reflects only an increase of a small number of additional books in the home by parents who were close to the Group 2/Group 3 cutoff point in 2000 and 2009. This is the sort of question that should be investigated before jumping to conclusions about achievement by social class in Finland.[40]

Similarly, the very dramatic increases in achievement by students in Germany across all social classes has received little media attention, although it does appear in the OECD report to Secretary Duncan as a phenomenon from which U.S. policymakers can learn. Germany had a PISA-reported increase of disadvantaged students in the 2000s, making the achievement in that decade of substantial gains for such students more interesting and possibly impressive. At the same time, the apparent concentration of German student performance gains in first- and second-generation immigrants, most from Russia and Eastern Europe, raises questions about whether school reforms were related to such gains or whether lessons learned in Germany from educating Russian and Eastern European immigrants are applicable to the U.S. context, where most low-scoring immigrants are from Mexico and come to the country with less educational background.

It is surprising that there has been so little focus on the reasons for the very large increases in U.S. mathematics scores in the past decade (and past 20 years), as measured by NAEP and confirmed to some extent by TIMSS. There is also evidence that students in some U.S. states, such as Massachusetts, score relatively high in mathematics when compared with students in similar post-industrial countries and even when compared with students in a "top-scoring" country, Canada. More attention should be paid by U.S. policymakers to the reasons for such relatively high performance in some states.

Part IX. Conclusion

Evidence-based policy has been a goal of American education policymakers for at least two decades. School reformers seek data about student knowledge and skills, hoping to use this information to improve schools. One category of such evidence, international test results, has seemingly permitted comparisons of student performance in the United States with that in other countries. Such comparisons have frequently been interpreted to show that American students perform poorly when compared to students internationally. From this, reformers conclude that U.S. public education is failing and that its failure imperils America's ability to compete with other nations economically.

This report, however, shows that such inferences are too glib. Comparative student performance on international tests should be interpreted with much greater care than policymakers typically give it. This care is essential for three reasons:

- First, because academic performance differences are produced by home and community as well as school influences, there is an achievement gap between the relative average performance of students from higher and lower social classes in every industrialized nation. Thus, for a valid assessment of how well American schools perform, policymakers should compare the performance of U.S. students with that of students in other countries who have been and are being shaped by approximately similar home and community environments. Because the distribution of students between social classes varies from country to country, differences in overall average scores between countries reflect, to varying extents, differences in school quality and differences in the degree of social inequality. Likewise, because the social class distribution also varies within the United States by state, comparisons of students in particular U.S. states where international tests are administered should also compare students in these states with students in other states and countries who have similar social class characteristics. Policymakers and school reformers may acknowledge these realities, but frequently proceed to ignore them in practice, denouncing relative U.S. international test performance with sweeping generalizations that make no attempt to compare students from similar social class positions.

  We have shown that U.S. student performance, in real terms and relative to other countries, improves considerably when we estimate average U.S. scores after adjusting for U.S. social class composition and for a lack of care in sampling disadvantaged students in particular. With these adjustments, U.S. scores would rank higher among OECD countries than commonly reported in reading—sixth best instead of 14th—and in mathematics—13th best instead of 25th.

- Second, to be useful for policy purposes, information about student performance should include how this performance is changing over time. It is not evident what lessons policymakers should draw from a country whose student performance is higher than that in the United States, if that country's student performance has been declining while U.S student performance has been improving. Policy implications become especially challenging if relative U.S. performance has been improving for some social class groups but deteriorating for others. Because U.S. policy is especially concerned with the performance of disadvantaged children, it would be wise to focus attention on trends over time of similar children in other countries, whether their overall national averages are higher or lower than the overall U.S. average. It makes little sense to hold up as successful models for the United States educational policies for lower social class students in countries where their performance is in sharp decline, even if trends in the average performance of all students in such countries obscures the performance of disadvantaged students.

This caution especially pertains to conventional attention to comparisons of the United States and higher-scoring Finland. Although Finland's average scores, and scores for the most-disadvantaged children, remain substantially higher than comparable scores in the United States, scores in the United States for disadvantaged children have been rising over time, while Finland's scores for comparable children have been declining. American policymakers should seek to understand these trends before assuming that U.S. education practice should imitate practice in Finland.

As well, U.S. trends for disadvantaged children's PISA achievement are much more favorable than U.S. trends for advantaged children. In both reading and math, disadvantaged children's scores have been improving while advantaged student's scores have been stagnant. U.S. policy discussion assumes that most of problems of the U.S. education system are concentrated in schools serving disadvantaged children. Trends in PISA scores suggest that the opposite may be the case.

- Third, different international and domestic tests sometimes seem to show similar trends, but sometimes seem quite inconsistent. These inconsistencies call into question conclusions drawn from any single assessment, and policymakers should attempt to understand the complex causes of these inconsistencies. Different tests that purport to reflect the performance of the same national cohort of students may sample students from different ages or grades. Different tests may also sample different aspects of the overall mathematics or reading curricula. Either or both types of considerations—differences in populations sampled or differences in curricular coverage—may explain the apparent inconsistencies in test results, but these factors have not been examined by policymakers. Without such examination, it is not possible to say whether the results of any particular international test are generalizable and can support policy conclusions.

In our comparisons of U.S. student performance on the PISA test with student performance in six other countries—three similar post-industrial economies (France, Germany, and the United Kingdom) and three countries whose students are "top scoring" (Canada, Finland, and Korea)—we conclude that, in reading:

- Higher social class (Group 5) U.S. students now perform as well as comparable social class students in all six comparison countries.
- Disadvantaged students perform better (in some cases, substantially better) than disadvantaged students in the three similar post-industrial countries, but substantially less well than disadvantaged students in the three top-scoring countries.
- The reading achievement gap between advantaged and disadvantaged students in the United States is smaller than the gap in the three similar post-industrial countries, but larger than the gap in the top-scoring countries.

We conclude that, in mathematics:

- U.S. students in all social classes perform relatively less well than in reading.
- Even so, disadvantaged students in the United States now do about the same or better than disadvantaged students in similar post-industrial countries, while advantaged students do much less well.
- U.S. students in all social classes perform less well than comparable social class students in the top-scoring countries.
- The mathematics achievement gap between advantaged and disadvantaged students in the United States is smaller than the gap in the three similar post-industrial countries, but mostly larger than the gap in the top-scoring countries.

Considering trends, the performance of disadvantaged U.S. students has improved between 2000 and 2009 in both reading and mathematics relative to the performance of disadvantaged students in five of our six comparison countries. This results both from the fact that disadvantaged students' average PISA scores in both tests declined or were unchanged in all comparison countries except Germany, while in the United States disadvantaged students' PISA scores have improved.

These comparisons suggest that much of the discussion in the United States that points to international test comparisons to contend that U.S. schools are "failing" should be more nuanced. Although claims about relative U.S. school failure often focus on disadvantaged students' performance, international data show that U.S. disadvantaged student performance has improved over the past decade in both mathematics and reading compared to similar social class students in all our comparison countries except Germany. TIMSS and NAEP data also show improvement for all social class groups in mathematics during the last decade. Should we consider these improvements a failure, particularly when the scores of disadvantaged students in all comparison countries but Germany have declined in this same period?

Data from both PISA and TIMSS suggest strongly that U.S. students—especially advantaged U.S. students—generally continue to do worse in mathematics, in contrast to their social class counterparts in comparison countries. Yet NAEP shows that mathematics is where academic improvement in U.S. schools has been the greatest—much greater than in reading. Thus, although the United States may have had and still has an incoherent, "mile wide and inch deep" mathematics curriculum, as identified in the most authoritative analysis of the first (1995) TIMSS test (Schmidt, McKnight, and Raizen 1997), math is apparently where U.S. students are making the largest gains across all social class groups.

To arrive at our conclusions, we made a number of methodological decisions. We have used a single measure of home literacy to define social class that we believe is the best measure. We have selected six countries based on their income levels or their status as high-scoring nations. We have estimated PISA scores for 2007 and, where possible, TIMSS scores for 2009, years in which these respective tests were not given. We have also transformed TIMSS scores to the PISA scale. In each of these, and in other cases, scholars and policymakers may choose different approaches. We believe our choices have been appropriate and have examined, where we could, the robustness of our results. We hope to inspire others researchers to pursue a similar inquiry.

We are most certain of this: To make judgments only on the basis of national average scores, on only one test, at only one point in time, without comparing trends on different tests that purport to measure the same thing, and without disaggregation by social class groups, is the worst possible choice. But, unfortunately, this is how most policymakers and analysts approach the field.

The most recent test for which an international database is presently available is PISA, administered in 2009. A database for TIMSS 2011 is scheduled for release in mid-January 2013. In December 2013, PISA will announce results and make data available from its 2012 test administration. Scholars will then be able to dig into TIMSS 2011 and PISA 2012 databases so they can place the publicly promoted average national results in proper context. The analyses we have presented in this report should caution policymakers to await understanding of this context before drawing conclusions about lessons from TIMSS or PISA assessments.

*— **Martin Carnoy** is Vida Jacks Professor of Education and Economics at Stanford University and a research associate of the Economic Policy Institute. He has written more than 30 books and 150 articles on political economy, educational issues, and labor economics. He holds an electrical engineering degree from Caltech and a Ph.D. in economics from the University of Chicago. Much of his work is comparative and international. His recent books include* Sustaining the New

Economy: Work, Family and Community in the Information Age, The Charter School Dust-Up *(coauthored with Richard Rothstein),* Vouchers and Public School Performance, Cuba's Academic Advantage, *and* The Low Achievement Trap.

— ***Richard Rothstein*** *is a research associate of the Economic Policy Institute and senior fellow of the Chief Justice Earl Warren Institute on Law and Social Policy at the University of California (Berkeley) School of Law. He is the author of* Grading Education: Getting Accountability Right *(Teachers College Press and EPI 2008) and* Class and Schools: Using Social, Economic and Educational Reform to Close the Black-White Achievement Gap *(Teachers College Press 2004). He is also the author of* The Way We Were? Myths and Realities of America's Student Achievement *(1998). Other recent books include* The Charter School Dust-Up: Examining the Evidence on Enrollment and Achievement *(co-authored in 2005); and* All Else Equal: Are Public and Private Schools Different? *(co-authored in 2003).*

— Queries and comments should be addressed to the authors at carnoy@stanford.edu and rrothstein@epi.org.

— We are enormously grateful to Tatiana Khavenson of the National Research University Higher School of Economics in Moscow for her invaluable research assistance, particularly in downloading and categorizing data from the PISA and TIMSS international databases. We thank David Grissmer, Edward Haertel, Helen Ladd, Sean Reardon, Richard Shavelson, and William Schmidt, who reviewed an early draft of this report. Their comments and suggestions were insightful and helpful, and resulted in an improved report. Larry Mishel, EPI president, read both an early and late draft, and made very helpful suggestions regarding both. Errors that remain after these reviews are solely the authors' responsibility. We frequently called on technical experts to answer specific questions, and we thank them for their willingness to assist, although they bear no responsibility for how we interpreted and used the information and data they provided: Yasin Afana, Dirk Hastedt, Barbara Malak, and Hans Wagemaker (the International Association for the Evaluation of International Achievement), Pierre Foy and Ina Mullis (the TIMSS and PIRLS International Study Center, Boston College), Richard Houang (Michigan State University), Miyako Ikeda and Sophie Vayssettes (the Organization for Economic Cooperation and Development), and Wolfram Schulz (the Australian Council for Educational Research). We are also grateful for extraordinary staff support we received from the Economic Policy Institute, especially our superb editor, Patrick Watson, publications director Lora Engdahl, research assistants Natalie Sabadish and Julia Cohen, graphic designer Dan Essrow, and editor Michael McCarthy. We also thank Eleanor Wang, formerly a graduate student at Stanford University, for research assistance at an early stage of this project.

Appendix A

Tables 3A and 3C of this report showed what the average 2009 PISA scores in math and reading would have been if the United States and the six comparison countries all had the average social class distribution of the three top-scoring countries. Tables 3B and 3D showed what the average 2009 PISA scores in math and reading would have been if the United States and the six comparison countries all had the average social class distribution of the three similar post-industrial countries. In general, with such standardization for social class, U.S. scores appear to be better than the actual average. Also in general, with standardization for the social class distribution of the top-scoring countries, the three similar post-industrial countries' scores appear to be better than their actual averages. And in general, with standardization for the social class distribution of the similar post-industrial countries, the three top-scoring countries' scores appear to be worse than their actual averages.

If the United States and the six comparison countries had the same social class distribution as the U.S. social class distribution (Table 2B, column (e) from the main report), and if average scores by social class were unchanged in the United States and in each of the comparison countries,

TABLE A1

Overall average scale scores, reading, for U.S. and six comparison countries, PISA 2009 (with standardization for U.S. social class distribution)

	TOP SCORING				SIMILAR POST-INDUSTRIAL				U.S.	U.S. VERSUS:	
	Canada	Finland	Korea	Average*	France	Germany	U.K.	Average*		Top-scoring average	Similar post-industrial average
National average reading score (from Table 1)	524	536	539	533	496	497	494	496	500	-33	+4
National average reading score, standardized for U.S. social class distribution	514	520	521	518	490	487	488	488	501	-18	+12
Difference between social class standardized reading scores and actual average reading scores	-10	-15	-18	-15	-6	-10	-6	-7	+1		

* Simple (unweighted) average of three countries

Source: Authors' analysis of OECD Program for International Student Assessment (PISA) 2009 database for each country

the overall national average scores in reading and math would appear as shown in **Tables A1** and **A2.**

With this standardization, reading and math scores are about the same as the nominal scores in France and the United Kingdom and, of course, in the United States. (U.S. scores, standardized by the U.S. social class distribution, should be identical to the nominal scores. Tables A1 and A2, however, show a difference of less than one point—rounded up to one point—in both reading and math. We do not know the reason for this discrepancy, but assume it is because not all test takers in the sample answered the BH question.) Social class standardized scores are worse than nominal scores in the other countries, with social class standardized scores substantially worse in Korea than nominal scores.

Appendix B

Books in the home (BH) and the Economic, Social, and Cultural Status (ESCS) indices

To check on the robustness of BH as a reasonable measure to capture social class groupings, we adjusted the average PISA scores in each BH category across our six key comparison countries by controlling for ESCS differences among individual students in each BH category in these countries. **Table B1** displays the average PISA scores *unadjusted* for ESCS differences, as reported in Tables 4 and 5 of the main text. **Table B2** displays the adjusted distributions. The scores for each social class group in each country in Table B2 are the average scores of students with that number of books in the home and whose ESCS scale scores are similar to students in the United States with that number of books.

TABLE A2

Overall average scale scores, mathematics, for U.S. and six comparison countries, PISA 2009 (with standardization for U.S. social class distribution)

	TOP SCORING				SIMILAR POST-INDUSTRIAL				U.S.	U.S. VERSUS:	
	Canada	Finland	Korea	Average*	France	Germany	U.K.	Average*		Top-scoring average	Similar post-industrial average
National average math score (from Table 1)	527	541	546	538	497	513	492	501	487	-50	-13
National average math score, standardized for U.S. social class distribution	517	529	524	524	492	502	487	494	488	-35	-5
Difference between social class standardized math scores and actual average reading scores	-10	-12	-22	-14	-5	-11	-5	-7	+1		

* Simple (unweighted) average of three countries

Source: Authors' analysis of OECD Program for International Student Assessment (PISA) 2009 database for each country

To estimate the adjusted scores in Table B2, we estimated regressions of individual student test scores within each of the BH social class categories for students in this group of countries as a function of their ESCS index value plus dummy variables for each of the countries, with the U.S. dummy left out as a reference. We conducted the regression analysis on each of the five sets of plausible values and average coefficients calculated from the five regressions. We used those average country regression coefficients to estimate the "adjusted" scale score for the U.S. reference dummy and for each country relative to the U.S. scale score. **Table B3** compares the unadjusted and adjusted BH results.

The unadjusted and adjusted reading scores are correlated 0.992 across BH categories and countries, and the unadjusted and adjusted math scores are correlated 0.990 across BH categories and countries.

Table B3 shows that the adjustment makes only a small difference in four of the six comparison countries and that in these countries, the two measures, ESCS and BH, are especially highly correlated. In Canada and Korea, however, there is a meaningful difference.

As we explain in the main text of this report, the PISA ESCS index relies, in part, on a count of physical articles in the home. Korean students report a much lower number of such articles than the other six countries in this group, France a somewhat lower number, Finland somewhat higher, and Canada much higher. This explains why adding the ESCS index to books in the home had the effect of raising Korean scores across all BH categories and lowering Canadian scores across all BH categories.

In the United States, including a control (by means of the ESCS index) for physical articles in the home adjusts the scores of more-advantaged students downward relative to less-advantaged students because of the larger differences in such physical articles among more- and less-advantaged students than their reported BH differences.

TABLE B1

Reading and mathematics scale scores by students' reported books in the home (BH), U.S. and six comparison countries, PISA 2009

Social class group (by BH)	Canada	Finland	Korea	France	Germany	U.K.	U.S.
Reading							
Group 1 (Lowest)	459	466	461	403	413	424	442
Group 2	492	495	501	458	455	455	471
Group 3	518	523	529	498	496	490	504
Group 4	543	552	546	533	523	522	529
Group 5	561	571	564	559	555	555	563
Group 6 (Highest)	567	572	581	573	551	562	563
National average	524	536	539	496	497	494	500
Mathematics							
Group 1 (Lowest)	471	490	452	413	433	435	434
Group 2	493	507	504	460	466	455	464
Group 3	521	528	531	498	509	487	491
Group 4	543	552	553	529	535	517	510
Group 5	560	570	579	562	571	547	548
Group 6 (Highest)	567	580	602	569	570	551	548
National average	527	541	546	497	513	492	487

Source: Authors' analysis of OECD Program for International Student Assessment (PISA) 2009 database for each country

However, we are not persuaded that having more articles in the home in Canada and fewer in Korea should be regarded as important in how students are advantaged or disadvantaged in ways that would affect their academic performance. Thus, we do not conclude that the data in Table B3 should cause us to reconsider our choice of BH as the measure by which to standardize academically relevant social class differences across countries.

The ESCS adjustment across BH categories suggests that using the BH categories does not pick up the entire social class effect on test score differences among U.S. students, although, as noted, we are not convinced that all elements in the ESCS index should be included in a social class index, or that they are relevant to academic achievement. Nonetheless, if we take into account ESCS in addition to BH, there would be a 113-point difference on the reading test between social class Groups 1 and 6, rather than a 121-point difference (see Table 4). And there would be an 87-point difference between social class Groups 2 and 5, rather than a 93-point difference. For mathematics, with use of ESCS in addition to BH, the difference between BH Group 1 and Group 6 would fall from 114 points to 107 points, and between BH Groups 2 and 5 from 84 to 78 points.

Mother's education, parents' education, ESCS, and books in the home as correlates of students' test scores

Another, similar, approach to checking how well the BH variable compares to alternative often-used variables—mother's education or highest parental education—as a proxy for student social class background is to correlate the BH measure with student test score and add mother's education or highest parental education (either mother's or father's, whichever is higher) as a second correlate to estimate how much the correlation changes.[41] If the change is small, it suggests that BH differentiates test scores by social class about the same as these other

TABLE B2

Reading and mathematics scale scores by students' reported books in the home (BH), adjusted for PISA Economic, Social, and Cultural Status (ESCS) index within BH category, U.S. and six comparison countries, PISA 2009

Social class group (by BH)	Canada	Finland	Korea	France	Germany	U.K.	U.S.
Reading							
Group 1 (Lowest)	454	461	470	410	415	422	442
Group 2	484	491	511	463	456	453	468
Group 3	509	518	539	504	495	487	500
Group 4	534	548	557	541	522	520	523
Group 5	552	567	575	565	551	553	556
Group 6 (Highest)	559	567	590	579	548	562	555
Mathematics							
Group 1 (Lowest)	464	484	462	419	434	432	433
Group 2	484	502	515	465	468	452	461
Group 3	511	522	542	505	508	484	486
Group 4	533	547	565	536	535	516	504
Group 5	550	565	590	567	567	544	539
Group 6 (Highest)	559	574	611	574	568	550	540

Source: Table B1, adjusted using regression analysis of individual test scores in each BH category controlling for individual ESCS index and country dummies

TABLE B3

Differences in reading and mathematics scale scores by students' reported books in the home (BH), adjusted and unadjusted for PISA Economic, Social, and Cultural Status (ESCS) index within BH category, U.S. and six comparison countries, PISA 2009

Social class group (by BH)	Canada	Finland	Korea	France	Germany	U.K.	U.S.
Reading							
Group 1 (Lowest)	-5	-5	8	6	2	-2	0
Group 2	-7	-4	10	6	1	-2	-2
Group 3	-9	-5	10	6	-1	-2	-4
Group 4	-9	-4	10	7	0	-1	-5
Group 5	-9	-4	11	5	-3	-3	-8
Group 6 (Highest)	-8	-5	9	6	-2	0	-8
Mathematics							
Group 1 (Lowest)	-6	-6	10	6	1	-2	0
Group 2	-9	-5	11	6	2	-2	-3
Group 3	-10	-6	11	7	-1	-3	-5
Group 4	-10	-5	12	8	0	-2	-6
Group 5	-10	-4	12	6	-3	-3	-8
Group 6 (Highest)	-8	-5	9	5	-2	0	-8

Source: Tables B1 and B2

TABLE B4

Correlation coefficients between student reading or mathematics score and various measures of student social class, for U.S. and six comparison countries, PISA 2009

Country	Sample size	Books in the home (BH)	BH + mother's education	BH + highest parental education	ESCS only
PISA 2009 reading score					
Canada	21,910	0.35	0.36	0.36	0.29
Finland	5,647	0.35	0.37	0.37	0.28
Korea	4,823	0.34	0.35	0.37	0.33
France	3,935	0.48	0.49	0.50	0.40
Germany	4,106	0.48	0.50	0.51	0.43
U.K.	10,984	0.47	0.47	0.47	0.36
U.S.	5,054	0.42	0.44	0.44	0.41
PISA 2009 mathematics score					
Canada	22,115	0.34	0.35	0.36	0.32
Finland	5,705	0.32	0.35	0.36	0.28
Korea	4,919	0.40	0.40	0.42	0.37
France	4,032	0.49	0.50	0.51	0.45
Germany	4,269	0.47	0.49	0.51	0.46
U.K.	11,211	0.47	0.47	0.47	0.40
U.S.	5,084	0.42	0.44	0.45	0.44

Source: Authors' analysis of OECD Program for International Student Assessment (PISA) 2009 database for each country

measures. If the change is large, it suggests that mother's education or highest parental education would likely differentiate test scores significantly differently from BH. In several of our sample countries (France, Germany, and the United Kingdom), about 5 to 9 percent fewer students answered the mother and father's education questions than the BH question, so we compare the correlations for the lower sample size.

Table B4 shows the results for the PISA reading and mathematics scores and **Table B5** for the TIMSS mathematics scores.

TABLE B5

Correlation coefficients between student mathematics score and various measures of student social class, for U.S. and select comparison countries, provinces, and states, TIMSS 2007

Country or province	Sample size A	Books in the home (BH)	BH + mother's education	BH + mother's education + articles in home	Sample size B	Books in the home	BH + highest parental education	BH + highest parental education + articles in home
British Columbia, Canada	4,053	0.28	0.28	0.34	4,072	0.28	0.31	0.36
Ontario, Canada	3,320	0.32	0.32	0.35	3,339	0.32	0.37	0.39
Quebec, Canada	3,826	0.31	0.32	0.34	3,839	0.31	0.33	0.35
Korea	4,219	0.41	0.41	0.45	4,221	0.41	0.46	0.49
U.S.	7,158	0.41	0.41	0.44	7,219	0.41	0.43	0.45
Massachuestts	1,857	0.46	0.46	0.48	1,860	0.46	0.47	0.49
Minnesota	1,745	0.34	0.35	0.39	1,745	0.34	0.39	0.42

Notes: Sample size A = sample size when including books in the home, mother's education, and articles in the home. Both BH and BH+ME. Sample size B = sample size when including books in the home, highest parental education, and articles in the home. Correlations shown for sample size A all use the same observations in sample A; correlations shown for sample size B all use the same observations in sample B.

Source: Authors' analysis of Trends in International Mathematics and Science Study (TIMSS) 2007 database (Boston College International Study Center) for each country, province, or state

Table B4 shows that adding either mother's education or highest level of parental education to the books-in-the-home variable adds very little to the correlation coefficient of books in the home with students' reading and mathematics scores. The table also shows that the PISA ESCS index is less correlated with reading and mathematics scores than books in the home. This does not mean that ESCS is a worse measure of social class; it only suggests that, if we think that the best social class variable should be most highly correlated with test scores, BH is a better predictor of students' academic achievement in reading and mathematics than is the PISA ESCS index.

Table B5 shows the correlations between TIMSS scores and various alternative measures of social class for those countries and provinces discussed in the main report. (Data are not available for England or Scotland.)

The table suggests that other variables add little to the correlation of BH with mathematics test scores. The biggest discrepancy is in the Ontario sample.

On the basis of these calculations, we conclude that, for these countries, adding other measures of social class or using other measures of social class to categorize social class groups does not improve significantly on our measure of using only books in the home.

Endnotes

1. PISA is sponsored by the Organization for Economic Cooperation and Development (OECD). See http://www.pisa.oecd.org/ *and* http://nces.ed.gov/surveys/pisa/. *PISA was administered to 15-year-olds in 2000, 2003, 2006, and 2009.*

2. TIMSS was administered by the International Association for the Evaluation of Educational Achievement (IEA) to 8th-graders in 1995, 1999, 2003, 2007, and 2011. See http://timss.bc.edu/ and http://nces.ed.gov/timss/. An international test of reading, the Progress in International Reading Literacy Study (PIRLS), was administered only to 4th-graders in 2001, 2006, and 2011. TIMSS was also administered to 4th-graders simultaneously with the 8th-grade administration. We do not analyze 4th-grade scores, either from PIRLS or from TIMSS, in this report.

3. NAEP is administered by the U.S. government, sporadically in many subjects on a national basis. Since 2003, however, Main NAEP 4th- and 8th-grade math and reading tests have been administered biannually in math and reading, with samples large enough to generate state-level results. Although state-level samples have been required by law only since 2003, many states voluntarily participated in this larger sampling as early as 1992. See http://nces.ed.gov/nationsreportcard/.

4. The average performance of students in Finland and Korea is a bit more than a third of a standard deviation better than that of average students in the United States, and the average performance of students in Canada is a bit less than a third of a standard deviation better than that of average students in the United States.

5. We also sometimes speak of "substantially higher (or lower)" interchangeably with "substantially better (or worse)," etc. And in the case of trends, we sometimes speak of scores that were "mostly unchanged," a phrase with identical meaning as "about the same."

Also making it difficult to interpret and compare the results from various assessments, each test has its own unique (and arbitrary) scale. In each case, however, when statisticians say that one country (or group) has an average score that is "significantly" better than that of a second country (or group), they mean that, given the distribution of test scores among sampled students in the two countries (groups), the probability is 95 percent or higher that the true average performance of students in the first country (or group) is better than the true average performance of students in the second country (or group) on that test (even if only a tiny bit better). When they say that one country (or group) has an average score that is "significantly" worse than that of a second country (or group), they mean that the probability is 95 percent or higher that the true average performance of students in the first country (or group) is lower than the true average performance of students in the second country (or group) on that test (even if only a tiny bit lower). And when they say that the average score in a country (or group) was "about the same" as the average score in a second country (or group), they mean that the true performance of the average

student in the first country (or group) would be neither significantly better nor significantly worse that the true performance of the average student in the second country (or group) on that test, in 95 percent of the times such a test were administered.

In PISA 2009, the standard error for each country's average score is not precisely the same, nor is the standard error for any particular country's average score necessarily identical to that of the OECD as a whole.

6. Eighteen scale points in most cases is equivalent to about 0.2 standard deviations. Policy experts generally consider an intervention that is 0.2 standard deviations or more to be an effective intervention; such an intervention, for example, would improve performance such that the typical participant would now perform better than about 57 percent of all participants performed prior to the intervention.

One reviewer of a draft of this report observed that for reasons discussed above, because of the well-known unreliability of a single test score, and because of differences in countries' alignment of their curricula with the PISA test (discussed below in Part VII), we should properly describe all differences that are less than 18 scale points as being "about the same." We do not disagree with this critique. However, we continue to describe differences of 8 scale points or more as being "better" or "worse" because the policy community has become so used to inappropriate descriptions of very small differences as meaningful that, were we to adopt the more appropriate cut-off of 18 rather than 8 scale points, readers might resist paying attention to these analyses. Perhaps at some future time, policymakers will be sufficiently comfortable with statistics that a report such as this could be written with an appropriate 0.2 standard deviation (in PISA, about 18 scale points) cut-off for definitions of "better" or "worse." This is not the situation today, however.

7. The difference between the text and the table is due to rounding (U.S. Group 1: 19.8 percent; U.S. Group 2: 17.6 percent). Throughout this report, subsequent apparent discrepancies of one point between whole numbers in text and table are also due to rounding. Thus, for example, throughout this report, we consider that a real change in PISA scores is one of 8 points or more. We consider a scale point difference of only 7.9 points to be "about the same," although it will appear, rounded, in a table as 8 points.

8. The gap appears to be the same in Germany and the United Kingdom only because of rounding. The difference in the gap between the United States and Germany is 7 scale points, and between the United States and the United Kingdom it is 8 scale points.

9. The instructions asked students to interpret a passage on the opposite page, when in fact the passage appeared on the previous page. PISA also notes that the confusion this error caused students might have affected the validity of their math scores as well (perhaps because U.S. test takers' overall confidence was shaken when they could not find the reading passage), but it regards the impact on math scores to be trivial. We are unable to make an independent judgment about how trivial this impact was, but note that U.S. math performance took an unusual and unexplained dip on PISA 2006, when scores were considerably lower than in both 2003 and 2009.

10. The change in the definition of the books-in-the-home categories between the 2000 survey and subsequent surveys could be an argument for using a different social class variable. However, the categories used for mother's and parents' education also changed between the 2000 and subsequent surveys. In Part IV we discuss the more complex issues that arise in using mother's education as a definer of social class, and also discuss the PISA social class index (the Index of Economic, Social, and Cultural Status, or ESCS). Also, as we discuss in Part IV, the use of BH as our social class variable makes it possible to compare PISA and TIMSS results by social class.

In 2000, unlike 2003 and subsequently, PISA was administered to different samples in math and in reading, and as a result the number of students in each BH category is slightly different for reading and math in 2000. For the estimates in Tables 8A and 13, we use a simple average of the reading BH group proportion and the math BH group proportion in 2000.

11. We estimate the interpolated scores by assuming that average scores increase linearly from category to category. For example, in 2000, the average reading score for U.S. students in the 11-50 books category was 480. We assume that this average score corresponded to students with the average number of books in the category—30 books (the midpoint from 11 to 50). The similar social class group, Group 2, in

the 2003, 2006, and 2009 PISA samples is defined as students with 11-25 books in the home, an average of 17.5 books. The next lowest social class category in 2000 was 1-10 books in the home, an average of five books, for which the average U.S. reading score was 431. We assume that U.S. students with 17.5 books would score lower than those with 30 books by the proportion (17.5-5)/(30-5) of the difference in test score (479 – 431, i.e., 24 points less, or 455). This is the average reading score we assign to the interpolated category of 11-25 books in the home (Group 2) in 2000. We make similar estimates for the interpolated categories, 26-100 books (Group 3), 101-200 books (Group 4), and 201-500 books (Group 5) for the 2000 PISA reading and math tests in the United States and each comparison country. For example, for the United States in 2000, the midpoints of the 51-100, 101-250, and 251-500 books categories are 75, 175, and 375, respectively. For the United States in 2003 and subsequently, the midpoints of Group 3 (26-100 books), Group 4 (101-200 books), and Group 5 (201-500 books) are 62.5, 150, and 350, respectively. Thus, the linear interpolation for adjusting the 2000 test scores are, for Group 3, (62.5-30)/(75-30) = 0.722; for Group 4, (150-75)/(175-75)= 0.75; and for Group 5, (350-175)/(375-175) = 0.875.

It is also necessary for analyses that follow in this report to estimate the distribution of 2000 test takers by BH groups as defined in 2003 and subsequently. For this purpose, reported distributions of Group 1 (10 books or fewer) and Group 6 (501 books or more) are comparable. The combined Groups 2 and 3 (11 to 100 books) and the combined Groups 4 and 5 (101 to 500 books) are also comparable. We estimate the sizes of Groups 2 and 3 by assuming that their weights, relative to each other, in 2000 were the same as in 2003, and we estimate the sizes of Groups 4 and 5 by assuming that their weights, relative to each other, in 2000 were the same as in 2003. To the extent that these assumptions are not perfect, some data in Figures B1, B2, C1, C2, D1, D2, and Tables 9 through 15 will deviate very slightly from their real values.

12. Table 9A shows that, in 2000, social class Group 6 students in France performed more poorly in reading than social class Group 5 students in that country. This, along with similar data for math (Table 11A), is the only case we have found where more advantaged students performed more poorly than the next lower social class group. Yet the data show that by 2009, social class Group 6 students had improved very substantially, and now, as in every other comparison in the seven countries under consideration, performed better than the next lower social class group. We have no explanation for this anomaly, but have rechecked the reported PISA data and confirmed that it is correct.

13. The numbers in Table 13 describing the actual average national scores for 2000 and 2009 are calculated by weighting the average scores by social class group by the actual proportion of that group in that year. These numbers differ slightly from PISA's reported national average scores (as shown in the bottom row of Tables 9A, 10a, 11A, and 12A), presumably because some students fail to answer the BH question when taking the test.

14. The sample size in each country varies as a fraction of the country's total 15-year-old population in school, and sampled subgroups in each country also vary as a fraction of their total in the 15-year-old population in school. The student weights reflect a number of adjustments, including bringing various sampled groups up to their proportion of the 15-year-old population, an adjustment for the fact that students in larger schools are more likely to be sampled than students in smaller schools, an adjustment for students missing from school at the time of testing, and adjustments for other sampling corrections.

15. Mother's educational attainment is also available for both PISA and TIMSS, but we chose BH rather than mother's educational attainment for reasons described in the text, above. Were we solely interested in within-country social class distinctions, we would probably consider mother's educational attainment to be a better proxy for social class status than BH. As Appendix B demonstrates, adding mother's educational attainment to BH to predict a country's PISA scores does not change the patterns we describe in this report.

16. We could, of course, have taken the continuous ESCS index and divided it into categories, establishing our own cut points to distinguish social classes. This would have added one additional complexity to our analysis, and raised questions about where we set the cut points. Certainly, however, the OECD's and IEA's cut points for BH are also arbitrary.

17. The PISA student questionnaire asks the question, "How many books are there in your home?" It then instructs the respondent, "There are usually about 15 books per foot of shelving. Do not include magazines, newspapers, or your schoolbooks." The respondent is then asked to tick one of the six categories we have listed. In PISA 2000, there was a seventh category of zero books. For our analysis of PISA 2000, we have combined the zero and 10 or fewer categories.

18. In Tables 3A and 3B, we showed how the average U.S. PISA reading score in 2009 would change if the U.S. had a social class distribution that was similar to those of top scoring or similar post-industrial countries, respectively. In Table 13 and Figures D1 and D2, we showed how changes in the social class distribution of countries over time, where social class is defined as BH groups, would influence countries' average PISA reading scores, absent any other changes. OECD makes similar estimates, using its ESCS index (OECD 2010b, Table II.3.2; OECD 2010c, Figure V.2.9). The trends it displays are similar to those we report for the seven countries on which we focus. However, OECD does not employ its ESCS index either to compare how students perform from similar backgrounds within countries from different ESCS backgrounds.

19. TIMSS was administered in 2007, PISA in 2006 and 2009. For this comparison, we construct an average of PISA scores by social class group in 2006 and 2009, with scores in 2006 given twice the weight. See below. PISA math scores for the United States dropped very steeply from 2000 to 2006, and then gained from 2006 to 2009. A comparison of TIMSS from 1999 to 2007 with PISA from 2000 to 2006 would show even greater inconsistency.

20. There was also a TIMSS administration in 1995, but because we are interested here primarily in comparison with PISA results beginning in 2000, we do not here examine TIMSS 1995.

21. For TIMSS, we consider a scale score to be "about the same" if it is 6 scale points or less, "better" if it is at least 7 points but no more than 16 points greater, and "substantially better" if it is 17 scale points or more. Seventeen scale points on the TIMSS in most cases is equivalent to about 0.2 standard deviations.

22. Because TIMSS has thus far released only national average scores in whole integers, it is possible, though unlikely, that a calculation from the database of the unrounded figure for TIMSS 2011 for the United States will result in a change from 1999 to "2009" of less than 7 scale points, which we would consider "about the same." In that case, we would consider that the trends from TIMSS and PISA for the United States in Table 14B corresponded.

23. PISA does not report comparable data for 2000.

24. This analysis cannot be extended to 2009 because Scotland did not participate in TIMSS 2011.

25. Table 16 constructs a result for PISA "2007" by averaging PISA 2006 results by social class group with PISA 2009 results for social class group, with PISA 2006 given twice the weight. Once TIMSS releases its international database for 2011, it will be possible to develop a companion table to test the validity of Table 16. The companion table will compare PISA 2009 with TIMSS "2009," constructed by averaging the performance by social class group in TIMSS 2007 and TIMSS 2011.

26. We convert TIMSS scores to the PISA scale by regressing PISA 2009 country average mathematics scores on TIMSS 2007 country average mathematics scores for 23 countries that took both tests. The correlation coefficient of the two tests is 0.93, and the equation used to simulate the PISA test score from the TIMSS is PISA score = 44.54 + 0.868 times the TIMSS score.

27. In 1998 in reading, and in 1996 and 2000 in math, the Main NAEP began offering accommodations to students with disabilities. Although the LTT claims to assess students on an unchanging set of skills, the test formats in reading and math were changed in 2004. In each of these cases, a test was administered to student samples in both the original and the new format (or test conditions); for these years the table displays an average of the two mean scores. (In each case, the national average results in both the old and new formats [or condition] were almost identical.) In subsequent tables and figures, where the discussion concerns only trends after the new format (or condition) was introduced, these table and figures use only the new format (or conditions).

28. In practice, NAEP does not seek 13-year-olds who are not in the 8th grade.

29. The Main NAEP was administered in 2007 and 2009; the table's report of 2008 Main NAEP average scores is an average of these two years.

30. As noted above, there are no PISA 2006 reading data for the United States because of an error in test administration. By interpolating a line from PISA 2003 to PISA 2009 for the United States, Figure E implies that U.S. PISA reading scores improved from 2003 to 2006. There is no basis for this inference. Reading performance for U.S. students on PISA could have declined further from 2003 to 2006, before rebounding to a higher level in 2009. Because this was the pattern in math for U.S. students from PISA 2003 to PISA 2006, and because this was also the pattern for the main NAEP in reading, this is at least a plausible alternative. There are no 2006 data for the LTT in reading, not because of an error in test administration but because no test was given. Again, the Figure E interpolation suggests that LTT reading scores improved from 2003 to 2006, and then again from 2006 to 2009. In view of the patterns in the other tests, a more plausible scenario might be that LTT average reading scores declined further from 2003 to 2006, and then rebounded more dramatically from 2006 to 2009.

31. For NAEP, with a standard deviation of approximately 34 in most cases, we consider scores to have improved if they gained about 3.5 points, and to have improved substantially if they gained about 7 points.

32. Certainly, not all African American students are disadvantaged, nor are all children whose mothers did not complete high school, nor are all children with few books in the home. But on average, racial minority status, low parental educational attainment, and indicators of little home literacy predict disadvantage.

33. In other words, if PISA were administered to 100 random samples of students, in 95 of those cases the results would be within about 7 points of the reported results.

34. For a review of scholarly literature on the impact of concentrated school poverty (peer effect) on student achievement, see Hanushek, Kain, Markman, and Rivkin, 2003.

35. National average scores in Table 25 differ slightly from the reported national average scores in Table 1 because Table 25 omits data for students who were not identified in schools by their schools' FRPL percentage and/or who did not answer the BH question. For 2008–2009, a year that corresponds to the PISA sample, the U.S. Department of Education's National Center for Education Statistics reports that 8 percent of all high school students attend schools where more than 75 percent of students participate in the FRPL program, up from 6 percent the previous year (NCES 2011, Figure 4). However, the department does not provide full data for 2008–2009, so Table 25 is based on complete data from 2007–2008.

36. This correspondence is available to interested researchers upon request.

37. These lists include only OECD countries. Sampled economies participating in PISA that are not nations (e.g., Shanghai) and non-OECD countries (e.g., Singapore, Chinese Taipei, and Liechtenstein) are not included.

38. Indeed, we are aware of no scholars who have investigated these inconsistencies.

39. Ruben Klein (2011) has analyzed the problems associated with sampling students in various grades at different times in the school year in various PISA test years, based on Brazilian PISA samples for 2000, 2003, 2006, and 2009. Klein shows that part of the mathematics gain, most of the science gain, and all of the reading gain from 2000 to 2009 resulted from increases in the grade level of students because of changes in the dates during the academic year when 15-16-year-old students were sampled for PISA.

40. We suggested in Part IV that an advantage of BH over ESCS is that BH is not a continuous measure and so it facilitates comparisons by social class groups as in this report. However, one disadvantage of a discontinuous measure like BH is the possibility that, in some instances, large numbers of sampled students could be clustered around group break points. We do not have any reason to believe that such clustering explains the Finnish trends described here but only suggest that it should be investigated.

41. See Chudgar, Luschei, and Fagioli (2012) for an application of this methodology to TIMSS scores.

References

Boston College International Study Center. TIMSS International Databases. Various years. 1995 (http://timss.bc.edu/timss1995i/Database.html), 1999 (http://timss.bc.edu/timss1999i/database.html), 2003 (http://timss.bc.edu/timss2003i/userguide.html), 2007 (http://timss.bc.edu/TIMSS2007/idb_ug.html).

Chudgar, A., T. F. Luschei, and L. P. Fagioli. 2012. *Constructing Socio-Economic Status Measures Using the Trends in International Mathematics and Science Study Data.* East Lansing: Michigan State University.

Dillon, Sam. 2010. "Many Nations Passing U.S. in Education, Expert Says." *New York Times*, March 10. http://www.nytimes.com/2010/03/10/education/10educ.html

Duncan, Arne. 2010. "Secretary Arne Duncan's Remarks at OECD's Release of the Program for International Student Assessment (PISA) 2009 Results." U.S. Department of Education, December 7. http://www.ed.gov/news/speeches/secretary-arne-duncans-remarks-oecds-release-program-international-student-assessment-

Duncan, Arne. 2012. "Statement by U.S. Secretary of Education Arne Duncan on the Release of the 2011 TIMSS and PIRLS Assessments." U.S. Department of Education, December 11. http://www.ed.gov/news/press-releases/statement-us-secretary-education-arne-duncan-release-2011-timss-and-pirls-assess

Fleischman, Howard L., Paul J. Hopstock, Marisa P. Pelczar, Brooke E. Shelley, and Holly Xie. 2010. *Highlights from PISA 2009: Performance of U.S. 15-Year-Old Students in Reading, Mathematics, and Science Literacy in an International Context* (NCES 2011-004). U.S. Department of Education, National Center for Education Statistics. Washington, D.C.: U.S. Government Printing Office. http://nces.ed.gov/pubs2011/2011004.pdf

Hanushek, Eric A., John F. Kain, Jacob M. Markman, and Steven G. Rivkin. 2003. "Does Peer Ability Affect Student Achievement?" *Journal of Applied Economics,* Vol. 18, No. 5, pp. 527–44.

Harmon, Maryellen, Teresa A. Smith, Michael O. Martin, Dana L. Kelly, Albert E. Beaton, Ina V.S. Mullis, Eugenio J. Gonzalez, and Graham Orpwood. 1997. *Performance Assessment in IEA's Third International Mathematics and Science Study (1995).* Amsterdam: International Association for the Evaluation of Educational Achievement.

Klein, Ruben. 2011. "A Reanalysis of PISA Results: Comparability Problems." *Ensaio: Avaliação e Políticas Públicas em Educação*, Vol. 19, No. 73, October/December. http://dx.doi.org/10.1590/S0104-40362011000500002

Martin, Michael O., and Ina V. S. Mullis. 2001. "TIMSS 1999 Benchmarking: An Overview." In M.O. Martin, K. Gregory, K. O'Connor, and S. Stemler, eds., *TIMSS 1999 Benchmarking Report.* Chestnut Hill, Mass.: TIMSS & PIRLS International Study Center, Boston College.

Mullis, Ina V. S., Michael O. Martin, Albert E. Beaton, Eugenio J. Gonzalez, Dana L. Kelly, and Teresa A. Smith. 1998. *Mathematics Achievement in Missouri and Oregon in an International Context: 1997 TIMSS Benchmarking*. Amsterdam: International Association for the Evaluation of Educational Achievement.

Mullis, Ina V. S., Michael O. Martin, Eugenio J. Gonzalez, Kelvin D. Gregory, Robert A. Garden, Kathleen M. O'Connor, Steven J. Chrostowski, and Teresa A. Smith. 2000. *TIMSS 1999 International Mathematics Report.* Boston, Mass.: Boston College, International Study Center.

Mullis, Ina V. S., Michael O. Martin, Eugenio Gonzalez, Kathleen M. O'Connor, Steven J. Chrostowski, Kelvin D. Gregory, Robert A. Garden, and Teresa A. Smith. 2001. *Mathematics Benchmarking Report, TIMSS 1999 Eighth Grade: Achievement for U.S. States and Districts in an International Context.* Chestnut Hill, Mass.: TIMSS & PIRLS International Study Center, Boston College.

Mullis, Ina V. S., Michael O. Martin, Teresa A. Smith, Robert A. Garden, Kelvin D. Gregory, Eugenio J. Gonzalez, Steven J. Chrostowski, and Kathleen M. O'Connor. 2003. *Assessment Frameworks and Specifications.* Chestnut Hill, Mass.: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., M. O. Martin, E. J. Gonzalez, and S. J. Chrostowski. 2004. *TIMSS 2003 International Mathematics Report Findings From IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades,* Chestnut

Hill, Mass.: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., M. O. Martin, G. J. Ruddock, C. Y. O'Sullivan, A. Arora, and E. Eberber. 2005. *TIMSS 2007 Assessment Frameworks.* Chestnut Hill, Mass.: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., M. O. Martin, and P. Foy (with J. F. Olson, C. Preuschoff, E. Erberber, A. Arora, and J. Galia). 2008. *TIMSS 2007 International Mathematics Report: Findings From IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades.* Chestnut Hill, Mass.: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., M. O. Martin, P. Foy, and A. Arora. 2012. *TIMSS 2011 International Results in Mathematics.* Chestnut Hill, Mass.: TIMSS & PIRLS International Study Center, Boston College.

National Center for Education Statistics (NCES) NAEP Data Explorer. http://nces.ed.gov/nationsreportcard/naepdata/

National Center for Education Statistics (NCES) online. *The Common Core of Data (CCD)*. U.S. Department of Education. http://nces.ed.gov/ccd/tables/2000_schoollunch_04.asp

National Center for Education Statistics (NCES). 2012. *The Condition of Education 2012.* U.S. Department of Education. http://nces.ed.gov/pubs2012/2012045.pdf

Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA). 2000 database. http://pisa2000.acer.edu.au/downloads.php

Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA). 2001. *PISA 2000: Knowledge and Skills for Life*. Paris: OECD.

Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA). 2003 database. http://pisa2003.acer.edu.au/downloads.php

Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA). 2004. *Learning for Tomorrow's World: First Results from PISA 2003.* Paris: OECD. http://www.oecd.org/dataoecd/1/60/34002216.pdf

Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA). 2006 database. http://pisa2006.acer.edu.au/downloads.php

Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA). 2007. *PISA 2006: Science Competencies for Tomorrow's World.* Paris: OECD.

Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA). 2009 database. http://pisa2009.acer.edu.au/downloads.php

Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA). 2010a. *PISA 2009 Results. What Students Know and Can Do: Student Performance in Reading, Mathematics, and Science (Volume I)*. Paris: OECD.

Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA). 2010b. *PISA 2009 Results: Overcoming Social Background – Equity in Learning Opportunities and Outcomes (Volume II).* Paris: OECD. http://dx.doi.org/10.1787/9789264091504-en

Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA). 2010c. *PISA 2009 Results: Learning Trends: Changes in Student Performance Since 2000 (Volume V).* Paris: OECD. http://dx.doi.org/10.1787/9789264091580-en

Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA). 2011. "Lessons From PISA for the United States."*Strong Performers and Successful Reformers in Education*. Paris: OECD. http://dx.doi.org/10.1787/9789264096660-en

Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA). 2012. *PISA 2009 Technical Report*. Paris: OECD.

Raudenbush, S. W., Y. F. Cheong, and R. P. Fotiu. 1996. "Social Inequality, Social Segregation, and Their Relationship to Reading Literacy in 22 Countries." In M. Binkley, K. Rust, and T. Williams, eds. *Reading Literacy in an International Perspective.* Washington, D.C.: National Center for Education Statistics, pp. 3–62.

Ripley, Amanda. 2011. "The World's Schoolmaster." *Atlantic*, July/August. http://www.theatlantic.com/magazine/archive/2011/07/the-world-8217-s-schoolmaster/8532/2/

Robelen, Erik. 2012a. "U.S. Math, Science Achievement Exceeds World Average." *Education Week,* Vol. 32, No. 15, December 11. http://www.edweek.org/ew/articles/2012/12/11/15timss.h32.html

Robelen, Erik. 2012b. "International Tests Spark Questions on Finland's Standing." *Curriculum Matters* (*Education Week* blog), December 20. http://blogs.edweek.org/edweek/curriculum/2012/12/educational_tourism_has_become.html

Schmidt, William H., Curtis C. McKnight, and Senta A. Raizen. 1997. *A Splintered Vision: An Investigation of U.S. Science and Mathematics Education.* Dordrecht, The Netherlands: Kluwer.

Schulz, Wolfram. 2005. "Measuring the Socio-economic Background of Students and Its Effect on Achievement in PISA 2000 and PISA 2003." Paper prepared for the Annual Meetings of the American Educational Research Association, San Francisco, Calif., April 7–11.

Stanat, P., D. Rauch, and M. Segeritz. 2010. "Schülerinnen und Schüler mit Migrationshintergrund." In E. Klieme, C. Artelt, J. Hartig, N. Jude, O. Köller, M. Prenzel, W. Schneider, and P. Stanat, eds. *PISA 2009. Bilanz nach einem Jahrzehnt.* Münster, Germany: Waxmann, pp. 200–230. http://www.pedocs.de/volltexte/2011/3536/pdf/Stanat_et_Al._Schuelerinnen_und_Schueler_D_A.pdf

This is a corrected version of a report initially posted on January 15, 2013. The corrections do not affect the report's conclusions. A description of the corrections, and the reasons for making them, can be found in "Response from Martin Carnoy and Richard Rothstein to OECD/PISA comments," posted online at http://www.epi.org/files/2013/EPI-Carnoy-Rothstein-Resp-to-Schleicher.pdf.